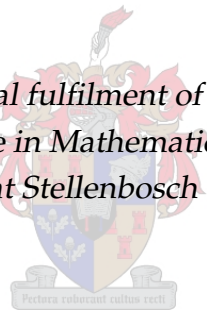
Analysis of partner turnover rate and the lifetime number of sexual partners in Cape Town using generalized linear models

by

Christianah Oyindamola Olojede



Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Mathematical Statistics in the Faculty of Science at Stellenbosch University

Department of Statistics and Actuarial Science,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor:

Prof. Wim Delva

December 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: .....
Christianah O. Olojede

Date: November 29, 2017 .....

Abstract

Analysis of partner turnover rate and the lifetime number of sexual partners in Cape Town using generalized linear models

Christianah O. Olojede

Department of Statistics and Actuarial Science,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc. (Mathematical Statistics)

June 2017

A large number of analyses have been carried out to investigate how sexually active people contracted human immunodeficiency virus (HIV) by using common indicators like the number of new sexual partners in a given year and the lifetime number of partners. In this study, the objective is to show that these are not always good indicators because what people report for these two indicators is not accurate nor consistent using generalized linear models such as Poisson and the negative binomial regression models. Generalized linear models are the types of models that allows for the distribution of the response variable to be non-normal. A cross-sectional, sexual behavioural survey was conducted in communities with a high prevalence of HIV in Cape Town, South Africa, in 2011 – 2012. We examined the effects of age and gender on the rate at which sexual partnerships are formed, using count data regression models. The age range of respondents was 16-40 years. The highest number of new sexual relationships formed in a year preceding the survey was 11 and the highest lifetime number of sexual partners was 15. A generalized linear regression model was used to examine the consistency between the reported number of new sexual partners formed in a year preceding the survey and the reported lifetime number of partners. We also assessed the predictive power of these two indicators for the respondent's HIV status. We found that these indicators are not consistent, and we conclude that they are not good indicators for predicting HIV status.

Keywords: Cross sectional survey, generalized linear model, HIV, negative binomial, Poisson, prevalence, regression, sexual partner.

Opsomming

Analise van Lewensmaat omset en die lewenslange aantal seksuele verhoudings in Kaapstad deur die gebruik van veralgemeende lineêre model

("Analysis of partner turnover rate and the lifetime number of sexual partners in Cape Town using generalized linear models")

Christianah O. Olojede

Departement Statistiek en Aktuariële Wetenskap,
Stellenbosch Universiteit,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc. (Wiskundige Statistiek)

Junie 2017

'n Hele aantal analises is reeds uitgevoer om ondersoek in te stel na hoe seksuele aktiewe persone menslike immuniteitsgebreksvirus (MIV) opdoen deur van die mees algemene indikators soos aantal nuwe seksuele metgeselle in 'n gegewe jaar asook die aantal lewenslange seksuele verhoudings te gebruik. In hierdie navorsing, is die doel om aan te toon dat dit nie altyd die beste indikators is om te gebruik nie omdat persone nie konsekwent of akkuraat die werklike aantal nuwe seksuele verhoudings in 'n jaar rapporteer nie deur veralgemeende lineêre model soos Poisson en negatief binomiaal regressie model le gebruik. 'n Veralgemeende lineêre model is die tipe model wat toelaat dat verspreiding van die responsveranderlike nie-normaal is. 'n Dwarsdeursnit opname oor seksuele gedrag is uitgevoer in gemeenskappe met hoë prevalensie van MIV in Kaapstad, Suid Afrika tussen 2011 en 2012. Die effek van ouderdom en geslag wat die vormingskoers van nuwe seksuele verhoudings beïnvloed, is ondersoek met behulp van kategoriese (tellings of frekwensies) regressie-modelle. Die ouderdomme van die respondente het gewissel tussen 16 en 40 jaar. Die maksimum aantal nuwe seksuele verhoudings gevorm in 'n jaar voor die opname was 11 en die maksimum aantal seksuele lewensmate waargeneem in die opname was 15. 'n Veralgemeende lineêre regressie-

model is gebruik om die konsekwentheid tussen die gerapporteerde aantal nuwe seksulele verhoudings in die voorafgaande jaar van die opname met die gerapporteerde aantal lewenslange seksuele verhoudings te bepaal. Die voorspelde onderskeidingsvermoe van hierdie twee indikators vir MIV status is ook geassesseer. Daar is gevind dat hierdie indikators nie konsekwent is nie en gevolglik nie wenslik is om MIV status te voorspel nie.

Sleutelwoorde: Dwarsdeursnit-opname, veralgemeende lineêre model, MIV, negatief binomiaal, Poisson , prevalensie, regressie, seksuele metgesel.

Acknowledgements

I would like to express my sincere gratitude to God Almighty, the maker of heaven and earth, who made this work a success. A popular saying goes thus: Only one person gives birth to a child but it takes a village to raise the child. This is true in my case, as this project would not have been a success without the wonderful people in my life.
I start by saying thank you to my parents, Mr and Mrs E.O. OLOJEDE; your support through the years can never be overemphasized. You are my inspiration and your words of encouragement keep me going. I say thank you once again for your teachings.
Many thanks to the South African Centre for Epidemiological Modelling and Analysis (SACEMA) for the financial, academic and moral support throughout this project. I am deeply appreciative to Prof. Wim Delva for his kind words of encouragement and supervision through it all, you are still the best.
I sincerely appreciate Dr Gavin Hitchcock and his wife, Rachel Hitchcock for their support during the editing process. A big thanks also goes to Roxanne Beauclair for helping me through the writing and editing phase. Thanks for your patience and understanding.
I say a big thank you to the love of my life, Oluwapamilerinayo. I would also like to appreciate Dr. Ogunleye Adeola, Dr. Olaoye Olufemi, Zinhle Mthombothi and my colleagues and friends at SACEMA. It is a pleasure to be part of the SACEMA family.

Dedications

I dedicate this work to the almighty and ever-knowing God, my parents, friends and family. You are the best.

Publications

A publication was extracted from this thesis. It is appended at the end of the thesis.

1. Investigating inconsistencies between the reported and predicted lifetime number of sexual partners in Cape Town, South Africa, published in the SACEMA Quarterly 2016.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Indicators used to predict the risk of having acquired HIV include the frequency of condom use, partner concurrency, intake of alcohol prior to sex, awareness of partner's exposure to HIV, number of new sexual partners in the past year, lifetime number of sexual partners, age, marital status, HIV prevalence in a community, male circumcision, multiple sexual partners, commercial or transactional sex, casual sex, and religion, among many others (Arora *et al.*, 2012; Kagaayi *et al.*, 2014). For the purpose of this study, we investigate two of the indicators – number of new sexual partners in the past year and the lifetime number of sexual partners to see how they vary by age and gender. Also, we check for inconsistencies in the reported and expected lifetime number of sexual partners, and whether these indicators are good predictors of HIV acquisition risk (Clumeck *et al.*, 2010; Friedland and Klein, 1987; Roberts *et al.*, 1986).

In our study, we considered heterosexual partnerships in three disadvantaged communities in Cape Town : Khayelitsha, Delft and Wallacedene.

1.2 The concept of partner turnover

This is the rate of sexual partner-change or new sexual partner acquisition per unit time. For the purpose of this study, we define sexual partner as an individual who shares intimate heterosexual moments with another. An individual can have a high or low rate of partner-change (discussed in detail in section 3.2). Individuals with high rate of partner-change are quick to acquire and transmit HIV infection (Anderson and May, 1988). A report by (Johnson *et al.*, 2001) in 2001 found a high rate of sexual partner-change among

individuals younger than 25 years in Britain. This rate is measured by the number of sexual partner per time. In our study, the data used categorizes the number of sexual partners into two; lifetime number of partners and the new number of sexual partners in a year preceding the survey. Lifetime number of sexual partners is the number of sexual partners an individual has had since sexual debut till the time the study was conducted. The number of sexual partners have been used in research to predict HIV risk (NE *et al.*, 1990; Quin *et al.*, 1986).
Our study used data from three communities in Cape Town, and the communities are described below.

1.3 The Khayelitsha population

Khayelitsha is an informal settlement found on the Cape flats near Cape Town International airport established in 1983. It consists mainly of the Xhosa people (Seekings, 2013) and a large part of the residents hail from the Eastern Cape. The population of Khayelitsha in 1996 was about 252,000, 329,000 in 2001 and 400,000 in 2011 (Seekings, 2013).

Just as in any multi–diverse society, there exist different races in Khayelitsha (see table 1.1). Its composition is made up of 99% black, 0.6% coloured (Here and henceforth in this thesis, "coloured" pertains to a particular racial group in South Africa, otherwise known as the Cape Coloured people) and infinitesimal number of white South Africans (SDI and GIS, 2013) as at 2011. The economic activities in Khayelitsha are unstable compared to other neighbouring towns and settlements.

Table 1.1: Racial distribution of residents in Khayelitsha (SDI and GIS, 2013)

Age	**Black African**		**Coloured**		**Asian**		**White**		**Other**	
(years)	Num	%	Num	%	Num	%	Num	%	Num	%
0 - 4	46246	12.0	277	12.0	26	9.6	25	7.6	199	8.0
5 - 14	62985	16.3	384	16.6	40	14.7	47	714.4	1904	4.2
15- 24	82552	21.4	418	18.1	61	22.4	58	17.7	712	28.8
25- 64	188245	48.7	1173	50.7	142	52.2	182	55.7	1450	58.6
$\geq$ 65	6330	1.6	63	2.7	3	1.1	15	4.6	11	0.4
Total	386358	100.0	2315	100.0	272	100.0	327	100.0		

1.3.1 The Delft population

Delft is a settlement established in 1989 and located near the city of Cape Town, South Africa. It is situated next to Khayelitsha and Capetown International Airport; it is a part of Tygerberg council area. The government of South Africa originally created Delft for low income coloured South Africans and most houses are subsidised by the government (Mongwe, 2002). The settlement is further divided into seven places, Delft Central, Eindhoven, Delft South, Voorbrug, Roosendal, The Hague and the new Symphony section (Delft, 2015). Delft is the first mixed race township in Cape Town (see table 1.2), consisting of both the coloured and black communities (Delft, 2015). In 2011, the Delft population was estimated to be 152,030 with 51.5% coloureds, 46.2% blacks and 2.2% others (including asians and whites). According to the 2011 census (SDI and GIS, 2013), the white and asian population residing in Delft together form less than 1% of the total population.

Table 1.2: Racial distribution of residents in Delft (SDI and GIS, 2013)

Age	**Black African**		**Coloured**		**Asian**		**White**		**Other**	
(years)	Num	%	Num	%	Num	%	Num	%	Num	%
0 - 4	9232	13.1	9679	12.4	49	9.4	20	11.2	275	9.9
5 - 14	12140	17.3	15092	19.3	86	16.4	17	9.6	252	9.0
15- 24	14126	20.1	16385	20.9	130	24.8	31	17.4	623	22.4
25- 64	34031	48.4	35895	45.9	249	47.5	104	58.4	1624	58.3
$\geq$ 65	6734	1.0	1228	1.6	10	1.9	6	3.4	13	0.5
Total	70263	100.0	78279	100.0	524	100.0	178	100.0	2787	100.0

1.3.2 The Wallacedene population

This is an informal settlement established during the 1980s and located in the eastern suburbs of Cape Town, South Africa. Wallacedene evolved when the South African government promoted the ending of Influx Control Act 68 of 1986, which halted the "pass laws" (Barry *et al.*, 2007). The land mass covered by Wallacedene is estimated as 0.54 square kilometres populated with about 21,000 inhabitants as at 2001. The population of Wallacedene in 2011 was 36, 583 with 10,392 total number of housing units (SDI and GIS, 2013). Most of the residents of Wallacedene are blacks just like Khayelitsha (see

table 1.3).

Table 1.3: Racial distribution of the people in Wallacedene (GIS, 2013)

Age	**Black African**		**Coloured**		**Asian**		**White**		**Other**	
(years)	Num	%	Num	%	Num	%	Num	%	Num	%
0 - 4	3870	13.2	778	13.3	5	8.8	12	10.3	132	11.8
5 - 14	4655	15.8	1224	20.9	6	10.5	18	15.4	65	5.8
15- 24	6434	21.9	1058	18.0	16	28.1	12	10.3	231	20.7
25- 64	14178	48.2	2702	46.0	248	49.1	66	56.4	676	60.5
$\geq$ 65	287	1.0	107	1.8	2	3.5	9	7.7	13	1.2
Total	29424	100.0	5869	100.0	57	100.0	117	100.0	1117	100.0

1.4 HIV in South Africa

HIV/AIDS in South Africa is a serious health issue. South Africa has one of the highest prevalence rates in the world and is among the countries with the largest number of persons infected with HIV/AIDS in the world (UNAIDS, 2014). Of the 36.7 million people infected with HIV/AIDS in the world, 17 million people live in Africa. In 2012, it was estimated that 12.2% of South Africa's total population of 50 million was HIV positive (Shisana *et al.*, 2005). In 2010, about 280,000 South Africans died as a result of HIV/AIDS. Despite the awareness created, it was evaluated that 350,000 to 500,000 new infections were being recorded annually (Navarro *et al.*, 2010). According to the studies carried out by the Human Science Research Council (HSRC), HIV/AIDS is highly prevalent in rural areas than urban areas. Illiteracy, lack of awareness about HIV/AIDS, and traditional beliefs and practises have greatly contributed to the spread of the disease. In South Africa, provinces with high HIV/AIDS prevalence are KwaZulu-Natal, Free State, Mpumalanga and the North West, while places with the lowest HIV/AIDS prevalence are the Northern Cape, Western Cape, and Limpopo (see Figure 1.1) (Shisana *et al.*, 2005). HIV/AIDS prevalence among the female gender is higher than among males in Eastern and Southern Africa (see Figure 1.2) (UNAIDS, 2016).

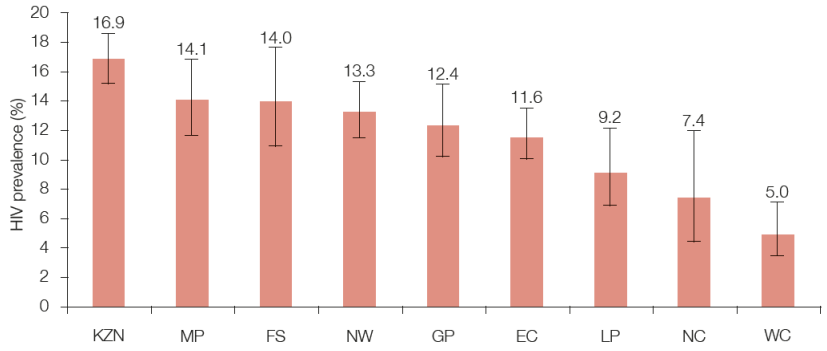


Figure 1.1: HIV prevalence by province in South Africa 2012, KZN – Kwazulu Natal, MP – Mpumalanga, FS – Free State, NW – North West, GP – Gauteng, EC – Eastern Cape, LP – Limpopo, NC – Northern Cape, WC – Western Cape (Shisana *et al.*, 2005)

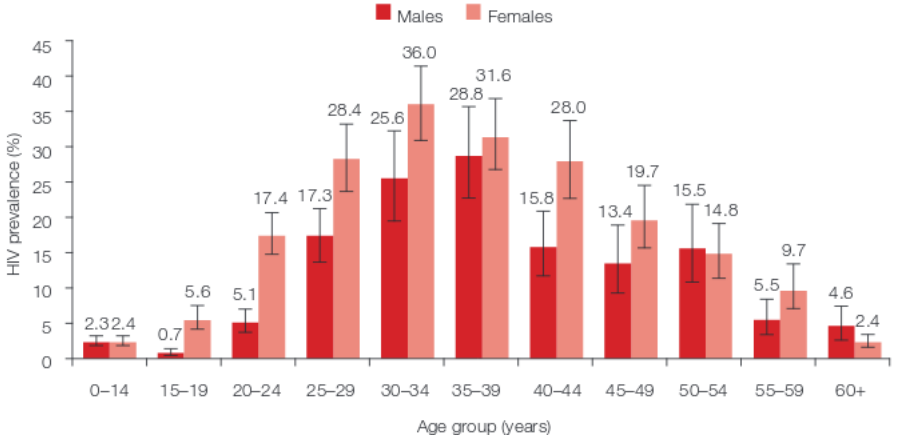


Figure 1.2: HIV prevalence by age and sex in South Africa 2012 (Shisana *et al.*, 2005)

1.4.1 Preventing HIV/AIDS

Use of condoms during sexual intercourse can greatly reduce the risk of contacting HIV although some people claim condom use makes sex unappealing by reducing sexual stimulation (Randolph *et al.*, 2007). Nowadays, we have female condoms that can be used by women. Only water-based lubricants should be encouraged because oil-based lubricants rip condoms apart.

Truvada drug is a type of nucleoside analog reverse transcriptase inhibitor (NRTI) used for the treatment of HIV infection. Some clinical trials have shown this drug to be protective against HIV for uninfected high-risk individuals if used daily – before and after exposure (Park, 2012). Medical physicians are mandated to test for hepatitis B infection

and if detected, the doctor should test the kidney functionalities before prescribing Truvada.

Sharing of sharp or body piercing objects between two or more individuals is known to increase the risk of HIV infection. Drug users are fond of sharing injections which can help transmit HIV virus from one person to another (Friedland and Klein, 1987). Among health practitioners in sub-Saharan Africa, after administering injection to a patient, the needle is disposed and the syringe is kept for further use.

HIV infection could be minimized by people notifying their sexual partners of their current HIV status before having sex / entering upon any conjugal relationship. Some couples are encouraged to go for testing before marriages and individuals living with HIV should be placed on anti-retroviral drugs to reduce the effect and replication of the virus in their bodies.

During the early pregnancy of an HIV positive patient, measures can be taken to prevent transmission from mother to child. It is estimated that 500,000 new-borns get infected with HIV in sub-Saharan Africa every year via mother-to-child transmission (MTCT) (Kak *et al.*, 2010).

Male circumcision also reduces the spread rate of HIV virus because the folded skin on an uncircumcised male body part can harbour the virus. HIV is less prevalent in regions that practise traditional circumcision compared to societies where men are not circumcised (WHO, 2007). According to the review of scholarly articles relating circumcision to HIV/AIDS prevalence, circumcised men are thrice less likely to contract HIV compared to uncircumcised men (WHO, 2007). About 3,274 uncircumcised male participants aged 18 to 24 were recruited for the South African Orange Farm trial. The outcome showed a 61% reduction against HIV infection. The trial was also carried out in Kenya and Uganda which shows 53% and 51% reduction in HIV acquisition for circumcised and uncircumcised men respectively (WHO, 2007).

Not engaging in risky sexual behaviours, celibacy, reduced number of sexual partners and no concurrent relationships, reduces the risk of HIV infection. Classifying various sexual practises into risky and safe is a bit challenging because of the fine lines between the two. Studies have shown that anal sexual practices carry higher risk than vaginal sexual practises due to the fact that mucous from the rectum differs from vaginal mucosa (PHAC, 2012).

1.4.2 Anti-retroviral drugs and therapy

Anti-retroviral drugs boost the immune system, prevent opportunistic infection, reduce HIV-related mortality and morbidity, inhibit mother to child transmission (MTCT) and suppress the viral load in an infected individual. The HIV virus can evolve and develop drug immunity, thus becoming virulent again under monotherapy. Therefore it is now universal practice to administer a combination therapy or anti-retroviral therapy (ART). The infected patients are expected to take the drug for the rest of their lives. Some of these ARV drugs have side-effects such as nausea, diarrhoea, skin rashes, sleep difficulties etc. A modern day ART known as Highly Active Anti-Retroviral Therapy (HAART) was developed and differentiated from the old ART by its name.

1.4.3 Factors that affect HIV acquisition risk

Rape and violence - women are the most vulnerable ones in this case. A woman could be raped by an already infected individual who in turn gets her infected.
In a case where formal education and employment are hard to get, individuals resort to getting money in exchange for sex. Money realized from transactional sex is a source of income. This could also affect the risk for the sex-worker of being infected and passing it on to multiple partners (Choudhry *et al.*, 2015; MacPherson *et al.*, 2012; Jewkes *et al.*, 2012).
High alcohol and drug intake causes people to practise risky sexual behaviours. For example, usage of condoms may be ignored when an individual is under the influence of alcohol (NIH, 2015).
Choice of partner is also an important factor. It is not only the number of sexual partners an individual has but also with whom they have sex (Roberts *et al.*, 1986; Roper *et al.*, 1993).
There are certain forms of sexual intercourse that increases HIV infection risk. As an illustration, after anal sex, an individual may decide to also go for vaginal sex. This takes the bacteria from the anus to the urethra, which increases the chance of getting infected (Wilton, 2014; Centers for Disease Control and Prevention, 2016; Winkelstein *et al.*, 1987).
Another important factor is migration. When infected individuals migrate to a community with low prevalence of HIV, there is a chance of fuelling an epidemic in that community especially if the migrants are sexually active (UNAIDS *et al.*, 2000; Lee *et al.*, 2012).
The socio-economic situation of the community could also be a factor. An individual's

HIV acquisition risk could be dictated by his/her social class in the society or even the socio-economic factor of the society, e.g low income, lack of education, high rate of unemployment, etc. (Hallman, 2009).
Religious beliefs could affect a person's HIV acquisition risk. Some religions encourage abstinence before marriage and faithfulness to marriage partner, so the people who practice the religion are less likely to have sexual intercourse until after marriage and are less likely to have sexual partners outside marriage (No, 2002).

1.5 Problem statement

There are many reasons why people have multiple sexual partners. This is usually measured in terms of the number of new sexual partners a person has had in the past year and the lifetime number of sexual partners. Can these indicators be trusted to predict HIV status well enough? There is need for more investigation of these indicators. By checking for inconsistencies in the reported and expected lifetime number of partners based on the reported number of new sexual partners in the last year, researchers can investigate whether these indicators are useful predictors of HIV status.

1.6 Research questions

In order to investigate the accuracy of the above mentioned indicators as good predictors of HIV acquisition risk, it is necessary to ask some questions that will lead us through the investigation. For the purpose of this study, three communities in Cape Town were used as case studies – Khayelitsha, Delft and Wallacedene. This thesis aims to address the following questions;

1. Is there variation in the rate of new relationships in terms of age and gender?
2. Is the reported lifetime number of partners consistent with the expected lifetime number of partners in this population?
3. After adjusting for the effect of age and gender, is there predictive power of the new number of sexual partners in the last year and the lifetime number of sexual partners on HIV status?

1.7 Significance of study

This study suggests that data and analysis conducted around the number of new sexual partners in a given year, and the lifetime number of sexual partners, needs to be re-visited with a healthy scepticism, and should not readily be taken at face value.

1.8 Definition of terms

1. Individuals – objects measured in a statistical problem
2. Data – measurements that have been recorded on individuals
3. Variables – characteristics measured on the individuals
4. Response – variable that measures the main outcome of the study
5. Sample – a part of the population being examined
6. Subjects – these are the individuals studied in the experiment
7. Questionnaire – research instrument consisting a series of questions for gathering information from respondents
8. Respondent – someone supplying information for a questionnaire

1.9 Thesis organization

The rest of the thesis is outlined as follows: Chapter two gives insight into the statistical methods used in this project. Chapter three discusses the background work that has been done in relation to partner turnover rate (PTOR). Chapter four presents the methods used in this study and the data collection procedure. Chapter five presents the results obtained from the analysis. Chapter six discusses the results obtained in chapter five and a detailed conclusion was drawn.

Chapter 2

Statistical methods

2.1 Overview

This chapter presents the statistical concepts used in this thesis. These are the statistical methods used to achieve the aims and objectives of this project. Here, we describe the model assumptions, specifications and derivations.

2.2 Cross-sectional study

Cross-sectional studies are population based studies, which do not follow individuals over time but are executed at a time point or in a little period of time. These studies are carried out to understand the prevalence of a disease or an exposure to a particular disease in a time period. It starts by selecting a sample from the target population and then obtaining data in order to group these individuals as having the disease or not having the disease. This type of study is used for descriptive purposes, when a population is being described with respect to an outcome of interest. It can also be used when we are interested in the prevalence of an outcome for a population or subgroup in a specific time-point.
Associations between an outcome of interest and its risk factors can also be investigated using cross-sectional studies. They indicate possibly existing associations and thus could be used in future research to create hypothesis (Levin, 2006; Alexander *et al.*, 2015).
Examples of cross-sectional study could be a census population conducted by the Census Bureau, say every 10 years, study on how much chocolate candy a student eats every week, a study of AIDS population in Africa, an experiment that tests whether or

not children who play video games are more violent than children who do not etc.
Advantages of cross-sectional study are, the cheapness of the experiment, it estimates the prevalence of the outcome of interest, no loss to follow-up as in longitudinal studies, helps in public health planning and many risk factors and outcomes of interest can be investigated. However, it is difficult to make causal inferences from a cross-sectional study because it does not indicate the sequence of events. Also, in the case of non-terminal diseases, there tend to be an under-representation of any risk factor that results in death (Levin, 2006).

2.3 Generalized linear models (GLMs)

Linear regression is used to describe the relationship between the mean of a response variable and some explanatory variables, assuming that the distribution of the response variable is normal (Agresti, 2015). Generalized linear models (GLM) accommodate the response distribution to be non-normal and they have three components: random, linear predictor and the link function components.
The random component is the response variable, say y, and its independent observations are $y_1, y_2, \ldots, y_n$. Here, independence is assumed, $E(Y) = \mu$, and the error variance constant σ^2 (McCullagh and Nelder, 1989).
The linear predictor, η, consists of parameters $\beta_1, \beta_2 \ldots, \beta_p$ where p is the number of explanatory variables $x_1, x_2, \ldots, x_n$, where n is the number of observations. This is represented as $\boldsymbol{X\beta}$, where $\boldsymbol{X}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is the parameter vector according to the following relation:

$$\eta = \sum_{1}^{n} x_p \beta_p$$

$$\eta = \boldsymbol{X\beta}.$$

It is generally assumed in GLM that the covariates enter the model through η (Winkelmann, 2013).
The link function g relates the expected value of the response variable to the linear predictor $\boldsymbol{X\beta}$ such that (Agresti, 2015; McCullagh and Nelder, 1989)

$$g\left[E(y)\right] = X\beta,$$

$$g(\mu) = \eta.$$

2.3.1 Likelihood functions of a GLM

Assuming each component of Y has an exponential family distribution, and is of the form,

$$f_Y(y;\theta,\phi) = \exp\{(y\theta - j(\theta))/i(\phi) + k(y,\phi)\}, \tag{2.3.1}$$

for some particular functions $a(.), b(.)$ and $c(.)$. This becomes an exponential family of canonical parameter θ if ϕ is known (Tutz, 2011; Agresti, 2015). For a Normal distribution, we have

$$f_Y(y;\theta,\phi) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\{-(y-\mu)^2/2\sigma^2\} \tag{2.3.2}$$

$$= \exp\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}(y^2/\sigma^2 + \log(2\pi\sigma^2))\}, \tag{2.3.3}$$

where $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, $c(y,\phi) = -\frac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}$.
The log-likelihood function, which is considered as a function of θ and ϕ with y given is represented as,

$$l(\theta,\phi;y) = \log f_Y(y;\theta,\phi).$$

The mean and variance are derived as shown below;

$$E\left(\frac{\delta l}{\delta\theta}\right) = 0 \tag{2.3.4}$$

$$E\left(\frac{\delta^2 l}{\delta\theta^2}\right) + E\left(\frac{\delta l}{\delta\theta}\right)^2 = 0. \tag{2.3.5}$$

From equation 2.3.1,

$$l(\theta;y) = \{y\theta - b(\theta)\}/a(\phi) + c(y,\phi),$$

hence

$$\frac{\delta l}{\delta\theta} = \{y - b'(\theta)\}/a(\phi) \tag{2.3.6}$$

and

$$\frac{\delta^2 l}{\delta\theta^2} = -b''(\theta)/a(\phi), \tag{2.3.7}$$

where $'$ and $''$ are differentiation with respect to θ. Therefore, from equations 2.3.6 and 2.3.7, it can be shown that

$$0 = E\left(\frac{\delta l}{\delta\theta}\right) = \{\mu - b'(\theta)\}/a(\phi), \tag{2.3.8}$$

then

$$E(Y) = \mu = b'(\theta).$$

From equations (2.3.5), (2.3.6) and (2.3.7),

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{\text{var}(Y)}{a^2(\phi)},$$

then

$$\text{var}(Y) = b''(\theta)a(\phi).$$

2.3.2 Deviance of a GLM

It is necessary to measure the discrepancy between the fitted model and the observation values when fitting a GLM. The statistic used in measuring this discrepancy is called the deviance, which is based on likelihood ratio statistic for comparing nested models. For a GLM with observations $\boldsymbol{y} = (y_1, \ldots, y_n)$, let $l(\boldsymbol{y}; \hat{\boldsymbol{\mu}}, \phi)$ represent the maximum likelihood function of the model where $\boldsymbol{y}^T = (y_1, \ldots, y_n)$ denotes the data and the fitted values, which is based on the maximum likelihood estimate is represented as $\hat{\boldsymbol{\mu}}^T = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$. The dispersion of observation is presented in the form $\phi_i = \phi a_i$, where a_i is known.
For every possible model, the best achievable log likelihood is $l(\boldsymbol{y}; \boldsymbol{y}, \phi)$, where $\hat{\mu} = y$. This is called the saturated model and it fits the data precisely as the observation parameters. Let $\theta(\hat{\mu}_i)$ denote the canonical parameter of the particular GLM of interest and $\theta(y_i)$ the canonical parameter of the saturated model. The deviance is then given as

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= -\phi 2\{l(\boldsymbol{y}; \hat{\boldsymbol{\mu}}, \phi) - l(\boldsymbol{y}; \boldsymbol{y}, \phi)\} \\ &= 2\sum_{i=1}^{n}\{y_i(\theta(y_i) - \theta(\hat{\mu}_i)) - (b(\theta(y_i)) - b(\theta(\hat{\mu}_i)))\}/a_i. \end{aligned}$$

The deviance of the model of interest is $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ and $D^+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/\phi$ is the saceled deviance, which compares the model of interest to the saturated model by $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \phi\lambda$. The deviance is mainly used for inferential comparisons of models and a great deviance indicates a poor fit (Tutz, 2011; Agresti, 2015).

2.4 Count data models

Some instances occur in which the response variables of interest is measured and recorded as non-negative integers. Number of occurrence of events or behaviour are measured in a particular time period. Examples of count data are the number of accidents that

occur in a town at a time point or particular period of time, number of individuals that died, number of customers in a bank at a time point, number of infected individuals in a population, etc.
Cameron and Trivedi used a data on the number of consultations with a Doctor to see the connection between insurance level and health care use (Cameron and Trivedi, 1986). Hall et al. studied the relationship between patenting, research and development expenditures by using the number of patents generated by firms (Hall *et al.*, 1986). Berko et al. examined weather related mortality by using the number of deaths attributed to weather (Berko, 2014). In these studies, count data were used to investigate relationship between variables.
Ordinary linear regression (OLS) is not suitable to model count data especially when the mean of the outcome is low. OLS produces biased standard errors (Gardner *et al.*, 1995). It may predict negative counts and the variance of the response variable may increase with the mean (Crawley, 2012). For these reasons stated above, regression methods like the Poisson regression, negative binomial regression, zero-inflated models, hurdle models, etc. have emerged to model count data.

2.5 Poisson distribution

Poisson process is a point process that is usually represented on a real line, which follows a stochastic process (Haight, 1967). It is used to model events, say, the arrival of clients in a bank e.t.c.
A Poisson distribution counts the Poisson process. This expresses the probability of the occurrence of an event in a specific interval of time. The average number of events in this interval is measured as λ. In the Poisson distribution, the values that may be observed do not necessarily have a finite upper limit. We give the probability distribution by

$$Pr\left(Y = y\right) = \frac{\exp^{-\lambda}\lambda^{y}}{y!}, \qquad y = 0, 1, 2, \ldots, n \tag{2.5.1}$$

where y denotes the dependent variables of some observed values n, and Y the Poisson distributed variable with rate parameter λ.
The random variable Y is Poisson distributed with λ and the length of time during which the events were measured, i.e

$$Pr\left(Y = y\right) = \frac{\exp^{-\lambda t}\lambda t^{y}}{y!}. \tag{2.5.2}$$

Equation (2.5.2) reduces to (2.5.1) if t is unity.
The moment generating function is given as

$$M(t) = e^{\lambda(e^t-1)}. \tag{2.5.3}$$

When $t = 0$, we have raw moments denoted by primes(')

$$\begin{aligned}
M'(t) &= e^{\lambda(e^t-1)} \cdot \lambda e^t \\
E(Y) &= M'(0) \\
&= e^{\lambda 0} \cdot \lambda e^0 \\
&= \lambda \\
M''(t) &= e^{\lambda(e^t-1)} \cdot \lambda e^t + \lambda e^t e^{\lambda(e^t-1)} \cdot \lambda e^t \\
E(Y^2) &= M''(0) \\
&= e^{\lambda 0} \cdot \lambda e^0 + \lambda e^0 e^{\lambda 0} \cdot \lambda e^0 \\
&= \lambda + \lambda^2 \\
Var(Y) &= E(Y^2) - [E(Y)]^2 \\
&= (\lambda + \lambda^2) - \lambda^2 \\
&= \lambda \\
Var(Y) &= E(Y) = \lambda.
\end{aligned} \tag{2.5.4}$$

This shows that the mean and the variance of the Poisson distribution is λ (Cameron and Trivedi, 2013; Haight, 1967; McCullagh and Nelder, 1989).

2.5.1 Poisson regression model

Poisson regression models are the standard models for count data because count data are non-negative integers and so the application of ordinary least square regression is not appropriate (Cameron and Trivedi, 2013). Poisson regression model happens to be a special case of the generalized linear model (GLM) with a log-link, which is the reason why it is also called a Log-linear model. It is derived from the Poisson distribution which is a bench-mark for count data (McCullagh and Nelder, 1989; Cameron and Trivedi, 2013; Winkelmann, 2013). Only the mean of the Poisson distribution determines the whole distribution as it is the only adjustable parameter as opposed the normal distribution which has two adjustable parameters, namely the mean and the variance. The response variable is modelled as having a Poisson distribution.
Cameron et al. used Poisson and negative binomial models to model the relationship between health care utilization and economic variables such as income and price by using

data from the Australian Health Survey from 1977 - 1978 (Cameron and Trivedi, 1986). Hausman et al. analysed the panel data on the number of the patents annually received by firms in the Unites States using Poisson regression. This was done to find the relationship between product innovation and research (Hausman *et al.*, 1984).
To model the relationship between the cost of usage and the demographic and economic characteristics of users, Ozuna et al. analysed data on the number of recreational boating trips to Lake Somoreveille in East Texas (Jaggia and Thosar, 1993).
Long used regression models to model the relationship between the amount of doctoral publications in the final years of PhD studies and number of articles by mentor, number of young children, mental status, etc. (Long and Freese, 2001).
To model the relationship between the number of defects per area in a manufacturing process and covariates like types of board surface, pad, panel and solder, Lambert used Zero-inflated Poisson regression model on data from soldering experiment (Lambert, 1992).
This type of regression models the natural logarithm of the expected value of the response Y, and it is given as

$$\log(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n. \tag{2.5.5}$$

The logarithm transformation of the mean ensures a positive value of response Y. Poisson regression does not have an additive error because the combination of an error on a linear scale and a log-linear mean function is not easy to interpret (Winkelmann, 2013).

2.5.2 Specifications of the Poisson regression model

There are assumptions that accompanies the Poisson regression models. The response (dependent) variable is denoted, y, and the explanatory (independent) variable(s) denoted as $\boldsymbol{x}$. The assumptions are discussed below (Winkelmann, 2013):

1. The conditional mean of y as a log-linear function of $\boldsymbol{x}$ and $\boldsymbol{\beta}$ is specified as

$$E(Y_i \mid \boldsymbol{x_i}) = \exp(\boldsymbol{x_i}\boldsymbol{\beta}) \qquad \text{with} \quad i = 1, \ldots, n \tag{2.5.6}$$

   where $\boldsymbol{x_i}$ is a $(1xp)$ vector of explanatory variables and $\boldsymbol{\beta}$ is a $(px1)$ vector of parameter in equation (2.5.6). An increase in $\boldsymbol{x\beta}$, which is important to acquire an unit increase in $E(Y \mid \boldsymbol{x})$ gets smaller as one moves away from zero, and it is specified by the exponential shape. Thus, the partial derivative is dependent on the value of $\boldsymbol{x\beta}$, where

$$\frac{\partial E(Y \mid \boldsymbol{x})}{\partial x} = \boldsymbol{\beta} \exp(\boldsymbol{x\beta}). \tag{2.5.7}$$

2. The conditional distribution of Y_i given $\boldsymbol{x_i}$ is given as

$$Y_i \mid \boldsymbol{x_i} \sim Po\left(\lambda_i\right). \tag{2.5.8}$$

Assumptions **1** and **2** combines to give the conditional probability law given as

$$Pr(Y_i = y \mid x_i) = \frac{\exp(-\exp(\boldsymbol{x_i\beta}))\exp(y\boldsymbol{x_i\beta})}{y!}, \quad y = 0, 1, 2, \ldots \tag{2.5.9}$$

Since there is only one parameter of the Poisson distribution, which determines both the mean and variance, then these two assumptions also determines the conditional variance of Y_i and it is given as

$$Var(Y_i \mid \boldsymbol{x_i}) = \exp(\boldsymbol{x_i\beta}). \tag{2.5.10}$$

Equation (2.5.10) is referred to as the **variance function**. The explanatory variables indirectly affect the response variable through the instantaneous occurrence rate of the process.

3. Poisson regression assumes that (y_i, x_i) are independently and identically distributed. This allows for a direct application of the maximum likelihood method to estimate the regression coefficients (Winkelmann, 2013).

2.5.3 Overdispersion

As shown in equation (2.5.4), the Poisson distribution has the conditional mean to be equal to the conditional variance and this is called equidispersion.
Poisson models are known to exhibit overdispersion and this occurs when the response variance is greater than the mean. If there exist an excess variation between the response counts, the existence of positive correlation between the responses and also when the distributional assumptions of the data are violated. This can cause underestimation of the standard errors, which makes a variable to be seen as a significant predictor when it is indeed not (Hilbe, 2012),

$$Var(Y) > E(Y), \tag{2.5.11}$$

$$Var(Y) < E(Y). \tag{2.5.12}$$

When the conditional variance exceeds the conditional mean, overdispersion occurs (2.5.11); but when the conditional mean exceeds the conditional variance, this is termed underdispersion (2.5.12). Overdispersion may occur due to unobserved heterogeneity.

Unobserved heterogeneity occurs if the explanatory variables are inadequate to account for the full amount of individual heterogeneity (Winkelmann, 2013). The magnitude of overdispersion or underdispersion can be measured by comparing the sample mean and variance of the response. Most of the time, count data are usually overdispersed than underdispersed (Cameron and Trivedi, 2013).

2.5.4 Modified Poisson regression model

This is a combination of the log Poisson regression model with robust variance estimation. This method is similar to a log binomial data, except that the model assumes that the response follows a Poisson distribution. It is used to rectify the problem that occurs when a Poisson regression is applied to binomial data and it yields overestimated error for the relative risk, alternatively called risk ratio (RR) (Zou, 2004). This is a problem of wide confidence interval especially when based on outcomes that are not rare. The modified Poisson regression approach often give valid confidence intervals. Zou proposed this model to model common binary data outcomes by incorporating a sandwich estimator into a log Poisson regression model to obtain robust error variance.
This approach is now a widespread substitute for the logistic regression model (Zou, 2004) and it's advantages lie in the fact that it estimates the relative risk directly rather than odds ratio provided by the logistic regression approach. Log-binomial regression often have convergence problems but is not the case with the modified Poisson regression (Wacholder, 1986)
Logistic regression is commonly used to model data with binary outcomes, with risk estimates reported as odd ratios, but Poisson regression (with robust sandwich variance estimator) can also be used to provide risk estimates and confidence intervals that are reasonable (Zou, 2004).

Table 2.1: Entries in a 2 by 2 table (Zou, 2004)

	x= 1 (event)	x=0 (no event)	Total
y=1 (exposed)	a	b	n_1= a +b
y=0 (unexposed)	c	d	n_0= c +d
			n= n_1 + n_0

Let us consider a situation whereby y_i $(i = 1, \ldots, n)$ is a dichotomous variable with a value 1 if exposed and 0 if unexposed. $\pi(y_i)$ is an underlying risk for subject i. We use

the logarithm link function to model $\pi(y_i)$ so as to obtain a positive estimate of $\pi(y_i)$. This gives

$$log[\pi(y_i)] = \alpha + \beta y_i.$$

We assume that x_i comes from a Poisson distribution, so the log-likelihood estimate can be written as

$$l(\alpha, \beta) = C. \sum_{i=1}^{n} [x_i(\alpha + \beta y_i) - \exp(\alpha + \beta y_i)],$$

where C is a constant and $\exp(\beta)$ is the relative risk (RR). The estimate of RR is given by

$$\hat{RR} = \exp(\hat{\beta})$$

and the estimated variance is written as

$$\hat{\text{var}}(\hat{RR}) = \frac{1}{a} + \frac{1}{c}.$$

The sandwich estimator corrects for error misspecification when the underlying distribution is binomial (Zou, 2004). Then the variance is estimated by

$$\hat{\text{var}}(\hat{RR}) = \frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_0}.$$

2.6 Negative Binomial regression model

The Poisson model was extended to overcome the problem of overdispersion in the data and this is called the negative binomial model (Patience and Osagie, 2014). The negative binomial model is equivalent to the Poisson regression model in many ways, in that it can be viewed as a Poisson-gamma mixture model but it has an extra parameter which accounts for dispersion (Hilbe, 2012).
It is assumed in the negative binomial regression that the Poisson parameter follows a gamma probability distribution. If the Poisson parameter for each observation i is written in the form

$$\lambda_i = \exp(\boldsymbol{X_i}\beta + \varepsilon_i),$$

where $\exp(\varepsilon_i)$ is an error term that is gamma distributed with mean 1 and variance α. This allows for variation between the mean and variance. Hilbe described that the negative binomial is not based on one derivation but can also be derived as a sequence of Bernoulli trials, could also be a type of inverse binomial distribution or a Polya-Eggenberger urn model (Hilbe, 2012). For the purpose of this study, we derive

the negative binomial as a Poisson-gamma mixture.
The probability density function of a negative binomial model can be derived from

$$f(y;\lambda,\mu) = \frac{\exp^{-\lambda_i\mu_i}(\lambda_i\mu_i)^{y_i}}{y_i!}. \tag{2.6.1}$$

Equation (2.6.1) is a Poisson model with gamma heterogeneity. Overdispersion is accomodated in the gamma mixture and the gamma noise has a mean of 1. Under conditional mean of y is $\lambda\mu$ rather than just λ under gamma heterogeneity,

$$f(y;x,\mu) = \int_0^\infty \frac{\exp^{-\lambda_i\mu_i}(\lambda_i\mu_i)^{y_i}}{y_i!} g(\mu_i)\partial\mu_i. \tag{2.6.2}$$

We derive the unconditional distribution of y from equation (2.6.2), where $\mu = \exp(\epsilon)$ and $\ln(\mu) = x\beta + \epsilon$. Let us assign a mean of 1 to the gamma distribution, then

$$f(y;x,\mu) = \int_0^\infty \frac{\exp^{-\lambda_i\mu_i}(\lambda_i\mu_i)^{y_i}}{y_i!} \frac{v^v}{\Gamma(v)} \mu_i^{v-1} \exp^{-v\mu_i} d\mu_i, \tag{2.6.3}$$

which gives,

$$f(y;x,\mu) = \frac{\lambda_i^{y_i}}{\Gamma(y_i+1)} \frac{v^v}{\Gamma(v)} \int_0^\infty \exp^{-(\lambda_i+v)\mu_i} \mu_i^{(y_i+v)-1} d\mu_i. \tag{2.6.4}$$

If we decide to move $\frac{\lambda_i^{y_i}}{\Gamma(y_i+1)} \frac{v^v}{\Gamma(v)} \frac{\Gamma(y_i+v)}{(\lambda_i+v)^{y_i+v}}$ to the left of the integral, then

$$\begin{aligned} f(y;x,\mu) &= \frac{\lambda_i^{y_i}}{\Gamma(y_i+1)} \frac{v^v}{\Gamma(v)} \Gamma(y_i+v) \left(\frac{v}{\lambda_i+v}\right)^v \frac{1}{v^v} \left(\frac{\lambda_i}{\lambda_i+v}\right)_i^y \frac{1}{\lambda_i^{y_i}} \\ &= \frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)} \left(\frac{v}{\lambda_i+v}\right)^v \left(\frac{\lambda_i}{\lambda_i+v}\right)^{y_i} \\ &= \frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)} \left(\frac{1}{1+\frac{\lambda_i}{v}}\right)^v \left(1-\frac{1}{1+\frac{\lambda_i}{v}}\right)^{y_i}, \end{aligned} \tag{2.6.5}$$

where :
Γ is the gamma function,
λ represents the mean of the distribution,
v is the dispersion parameter,
y is the dependent variable or response.

We invert the overdispersion parameter, v, to give us α, and equate λ and μ. This results in the negative binomial probability mass function given below:

$$f(y;\mu,\alpha) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(1-\frac{1}{1+\alpha\mu_i}\right)^{y_i}. \tag{2.6.6}$$

It follows that $\Gamma(y+1) = y!$, $\Gamma(y+\frac{1}{\alpha}-1) = (y+\frac{1}{\alpha})!$, and $\Gamma(\frac{1}{\alpha}) = (\frac{1}{\alpha}-1)!$
Then

$$f(y;\mu,\alpha) = \binom{y_i+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}. \tag{2.6.7}$$

Therefore, equation (2.6.7) is the probability mass function of a negative binomial distribution (Hilbe, 2012).
The mean and variance of the negative binomial distribution are given below;

$$E(Y) = \mu_i,$$
$$Var(Y) = \mu_i + \alpha\mu_i^2.$$

The log-likelihood function of the negative binomial distribution is given as

$$\begin{aligned}\mathcal{L}(\mu;y,\alpha) = \sum_{i=1}^{n} y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i+\frac{1}{\alpha}\right)\\ - \ln\Gamma\left(y_i+1\right) - \ln\Gamma\left(\frac{1}{\alpha}\right).\end{aligned} \tag{2.6.8}$$

The negative binomial probability density function can also be described as the probability of having y failures before rth success in a sequence of Bernoulli trials. Equation (2.6.7) becomes;

$$\begin{aligned} f(y;p,r) &= \binom{y_i+r-1}{r-1} p_i^r (1-p_i)^{y_i}, \\ &= \frac{(y_i+r-1)!}{y_i!(r-1)!} p_i^r (1-p_i)^{y_i}, \end{aligned} \tag{2.6.9}$$

where $\alpha = \frac{1}{r}$ (as in 2.6.7), p is the probability of r successes and y is the number of failures before the rth success.

2.7 Splines

2.7.1 Piecewise polynomials

Let us assume we have a given set of data points $(x_1,y_1),\ldots,(x_m,y_m)$, where

$$a = x_1 < x_2 < \ldots < x_{m-1} < x_m = b,$$

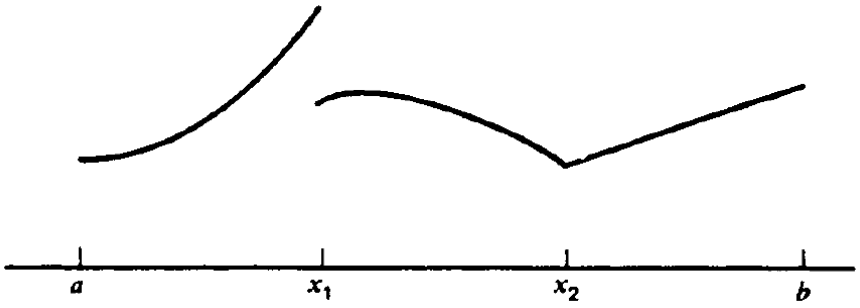


Figure 2.1: A quadratic piecewise polynomial (Schumaker, L., 1981)

where x_0 and x_m are the boundary or end knots (Shikin E.V. and Plis A.I., 1995). Interval $[a, b]$ is partitioned into x_m subintervals and we use a low degree polynomial to approximate function $f(x)$ on each subinterval.

$$S(x) = f(x_i) + (x - x_i)\frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}, \quad \text{if} \quad x \in [x_i, x_{1+1}]$$

Schumaker (Schumaker, L., 1981) gave a diagrammatic illustration of a quadratic piecewise polynomial of order 3 with 2 knots in Figure 2.1. Polynomial spline functions are not essentially smooth and they can also be discontinuous as in Figure 2.1 (Schumaker, L., 1981). In practical applications, we would prefer a relatively smooth function, which are called splines, because polynomials are inadequate to approximate functions which arises from physical world and not the mathematical world. In the physical world, behaviour in one region could be unrelated to behaviour in another region, thereby giving rise to their disjoint nature, which can be accommodated by spline functions because they have a piecewise nature (Wold, 1974).

2.7.2 Types of splines

Splines are smoothly connected piecewise polynomial approximations. They are connected at the polynomial pieces (knots) $x_i's$, $i = 1, \ldots, m$ with different continuity conditions. We have different types of splines, they are: linear, quadratic and cubic splines as shown below.

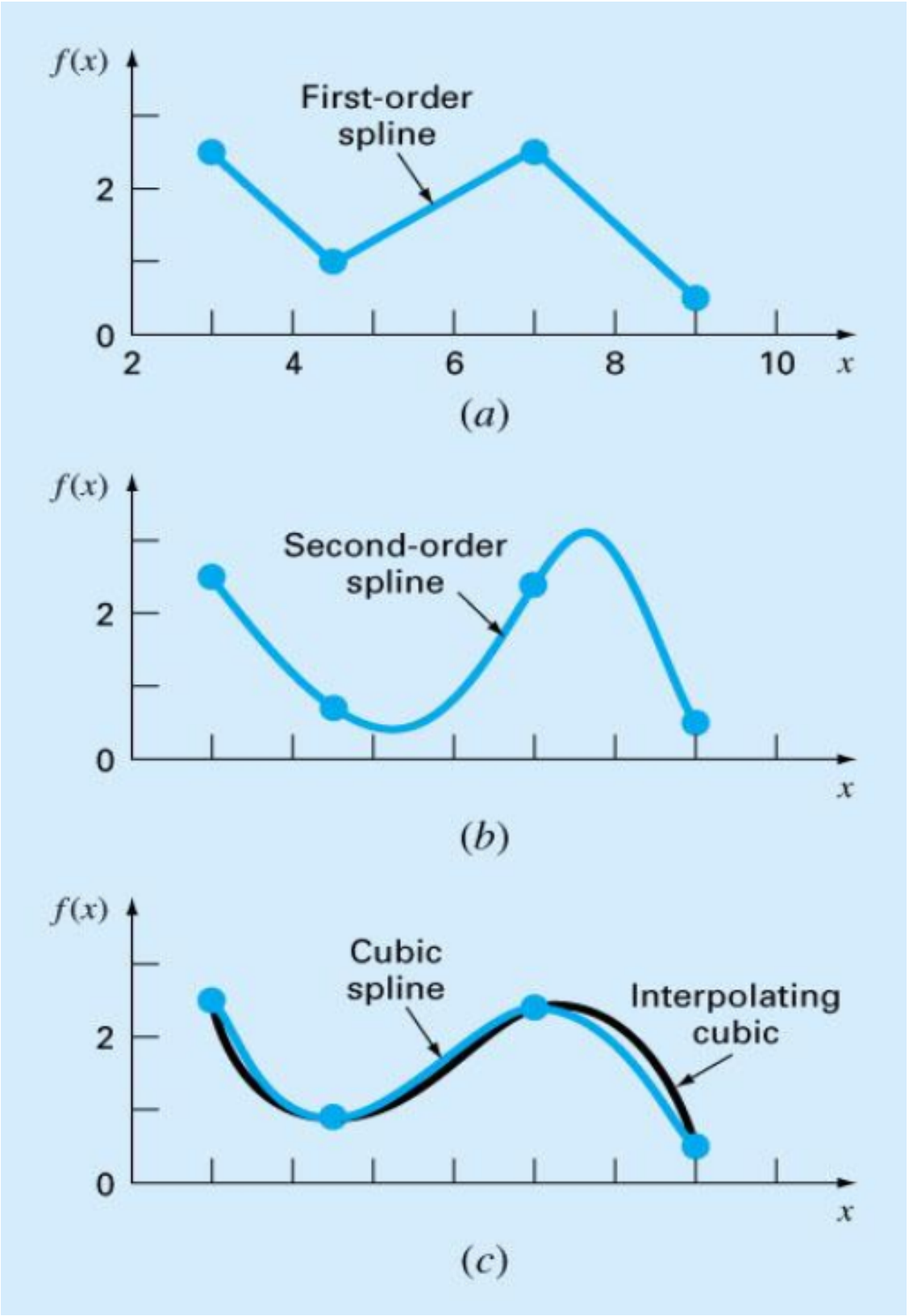


Figure 2.2: (a) shows an example of what a linear spline looks like. It is apparent that the derivatives are not continuous and not smooth (b) shows the example of a quadratic spline where the first derivatives are continuous (c) illustrates a cubic spline with both the first and second derivates continuous and it is the smoothest of the three (Huang, 2012).

1. Linear spline: $S_i(x) = a_i x + b_i, \quad \text{for} \quad x \in [x_i, x_{i+1}]$.
2. Quadratic spline: $S_i(x) = a_i x^2 + b_i x + c_i, \quad \text{for} \quad x \in [x_i, x_{i+1}], \quad i = 1, 2, \ldots, n-1$.
3. Cubic spline is detailed in section 2.7.3.

For the purpose of this project, we focus on using the cubic spline in order to achieve smoothness.

2.7.3 The cubic spline

This is a third-order polynomial spline that passes through a set of m control points. The usual starting point in studying spline function is the cubic spline (Ahlberg *et al.*, 1967). They are the most popular spline functions. High-degree polynomial interpolations are known for oscillatory behaviour, but cubic splines possess stability (Atkinson, 1989). We define a function S(x) as a cubic spline function, if when defined on a grid $\mho$ is

1. a cubic polynomial

$$S(x) = S_i(x) = a_0^{(i)} + a_1^{(i)}(x - x_i) + a_2^{(i)}(x - x_i)^2 + a_3^{(i)}(x - x_i)^3$$

   on everyone of the partial interval $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, m-1$.

2. in possession of a second derivative that is continuous on the interval $[a, b]$, and
3. satisfying the conditions

$$S(x_i) = y_i, \; i = 0, 1, \ldots, m$$

where n is the total number of the partial intervals, $m-1$ is the number of the inner knots (Shikin E.V. and Plis A.I., 1995).
A cubic spline with K knots can be represented as follows:

$$S(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \gamma_j (x - \kappa_j)_+^3 .$$

We have different types of cubic splines which are: natural cubic spline, end slope spline, periodic spline, not-a-knot spline, etc.
Since polynomial fits are known with the problem of inconsistency near the boundaries, splines could even compound this problem the more. For the purpose of this project, we use the natural cubic spline method.

2.7.4 Natural cubic spline

This problem of inconsistency has been resolved by the addition of more constraints to the boundary knots, thereby forcing it to be linear. Two degrees of freedom are freed on each boundary (making it four altogether), and since we have less information near the boundaries, we can afford to restrict them to be linear. It is also called the restricted cubic splines.
A natural spline that has k knots has k degrees of freedom and a natural spline has $n + k - 4$ degrees of freedom. It is of the form:

$$S(x) = \beta_0 + \beta_1 x + \sum_{j=1}^{k} \gamma_j \left(x - \kappa_j\right)_+^3,$$

subject to restrictions

$$\sum \gamma_j = 0 \text{ and } \sum \gamma_j \kappa_j = 0.$$

and this leaves us with k parameters (Rodrıguez, 2001; Hastie *et al.*, 2009).
Harrell states that the number of knots that is suitable for a large data set is four as it is a good compromise between the flexibility and the inaccuracy, which is caused by fitting a small sample (Harrell, 2015). We use the cubic spline in our analysis to achieve a higher degree of smoothness due to the fact that both the first and second derivatives are continuous at the knots and the natural cubic spline to prevent both ends of our graphs from distortion.

2.8 Synthetic cohort approach

A group of people who experienced a particular (common) event in a time period, usually a year of birth, is a cohort. When there is an interest in measuring experience from an event or behaviour of a cohort, it is sometimes impossible to wait until all the members of the cohort have had their experience of that particular event to get the needed information. For instance, in the study of divorce, synthetic cohort approach measures estimate the incidence of divorce among the cohorts that are currently marrying. This is observed at a time point but it focuses on future experience. This means that the cohort study was not actually carried out. The disadvantage here is that the method assumes that age-specific divorce rates will be constant into the future (Halli and Rao, 1992).
Attanasio used the synthetic cohort approach to examine financial asset accumulation in the United States (Attanasio, 1993),

2.9 Bootstrap

Bootstrap is a nonparametric simulation method, which is data-based and it is used to compute confidence intervals and make inferences (Harrell, 2015). This is not the same "bootstrap" used in computer science. It is a data re-sampling method which uses the information from the sample instead of specifying the data-generating process. No assumption is made about the distributions or the true values of the parameter (Efron and Tibshirani, 1994). One good thing about the bootstrap is that the approximations converge faster for some statistics compared to the approximations based on asymptotic theory. Bootstrap can be used when the asymptotic sampling distribution is too difficult to derive or too time-consuming or too expensive. It can be used to produce consistent approximations for some estimators like the mean, median, standard deviation, confidence bounds etc. For the purpose of this study, this method was employed to construct confidence bounds around the bias curves.
Suppose we have a population with a distribution function F and a random sample size n which gives $X_1, X_2, \ldots, X_n$ was drawn. Assuming we want to estimate the mean μ of the population, we have:

$$\mu = \int xF(x).$$

The empirical distribution function is given as:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x).$$

The bootstrap estimator of the population mean μ is the sample mean which is given as:

$$\bar{X} = \int x d\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A bootstrap sample is defined as a random sample size n drawn from the empirical distribution function $\hat{F}$, for example,

$$X^* = (X_1^*, X_2^*, \ldots X_n^*).$$

2.10 Cross validation

When we are carrying out a regression analysis, we are concerned with the error measurement. A type of quantity that measures the accuracy with which a model predicts the response value of a future observation is called the prediction error. A tool for estimating the prediction error is cross-validation. The expected squared difference between a response value and its prediction from a regression model is the prediction

error which is represented as:

$$\mathrm{PE} = \mathrm{E}\,(y - \hat{y})^2\,,$$

where E is the expectation and is the repeated sampling from the original population. In classification problems, the probability of a misspecification is called the prediction error and it is represented as :

$$\mathrm{PE} = \mathrm{Prob}\,(\hat{y} \neq y)\,.$$

Cross-validation is a process whereby a data set is split into two parts, namely, training and test data. The training data set is used to train a regression model and its accuracy when applied to the test data set gives the error estimate. The model is fitted to the training data set and we predict the responses from the observations by using the fitted model.

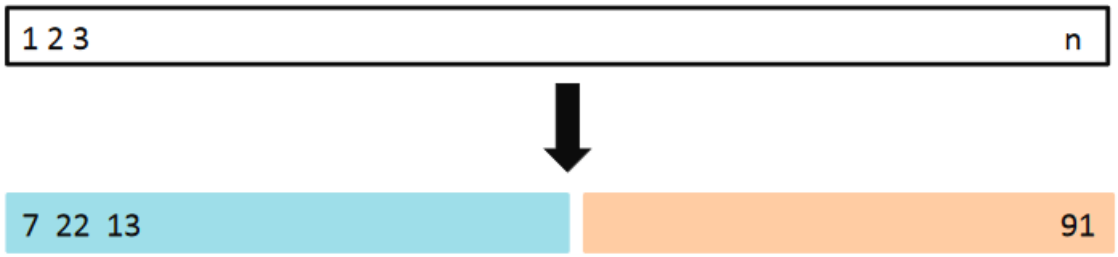


Figure 2.3: Source: Cross-validation and the bootstrap (Efron and Tibshirani, 1995)

Figure 2.3 shows the splitting process (random) of the data set (1 to n) ; the left part $(7, 22, 13, \ldots)$ is the training set and the right part $(\ldots, 91)$ is the test set. The data is split because we cannot use the same data to train the model and also test it. This is done so we could get a more realistic estimate of the prediction error.
In our analysis, we focus on a particular type of cross-validation called leave-one-out cross-validation (LOOCV). Here, just one data point is used as the test data. Then a model is built on the remaining data set which in this case is the training data, and the error is evaluated on the single data point removed from the data. We obtain the prediction error by repeating the procedure for each of the training data points left.
Assume that we split the data into k parts. Take $\hat{y}_i^{-k(i)}$ as the fitted value for observation i which is computed with the $k(i)$th part of the data removed. In the LOOCV, $k = n$. The cross validation estimate of the prediction error is then given by (Efron and Tibshirani,

1994):

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{-k(i)}\right)^2.$$

This is not the ideal method to use when we have a large data set because it is computationally expensive but it works well for smaller data sets.

Chapter 3

Literature review

3.1 Introduction

Sexual dissatisfaction, distance between partners, unemployment, infidelity, substance use and abuse, youthful exuberance (adventure), gender-based violence, demographic turnover, place of birth, low educational level, low income, religion, age at first intercourse, marriage and exposure, are some of the factors that contribute to partner turnover (Borgdorff *et al.*, 1994; Tanfer and Schoorl, 1992; Cooper *et al.*, 2012). The next section reviews partner turnover rate.

3.2 Partner turnover rate (PTOR)

Partner Turnover Rate (PTOR) can be defined as the rate at which new (sexual) relationships are formed between individuals. In a study where data from the National Survey of Unmarried Women was used, Tanfer et al. (Tanfer and Schoorl, 1992) found that partner turnover rate is tied to experience and exposure as older women had more sexual relationships than their younger counterparts. However, this cannot be generalized to all women as the study only included the never–married women. Apart from the fact that self-reported data may not be accurate, the study also assumed that the women were monogamous during a long–term relationship. This may or may not be the case for all subjects in the study. In our subsequent sections, we will expatiate on the reasons for forming new relationships.
PTOR can be studied in two categories: individuals who have the propensity to have many new sexual partners within a short period say a year (high PTOR), and individuals who tend to have fewer new sexual partners in a short period of time (low PTOR).

Johnson et al. (Johnson *et al.*, 2009) classified these two groups as the high and low risk groups respectively. For example, individual A had ten new number of sexual partners in the last year and individual B had two new sexual partners in the last year. In this illustration, individual A has a high PTOR and individual B has a low PTOR.

An article by USAID in 2010 links an individuals's risk of acquiring HIV directly to the number of sexual partners over time and their partner's behaviour. They used the illustration of an individual with only one sexual partner, but the partner is connected to a wide sexual network through concurrent relationships (sexual partnerships that overlay in time). The individual is said to have a higher risk of contracting HIV. Having more than one sexual partner irrespective of the pattern, either concurrent or multiple, still increases an individual's risk of HIV acquisition (Epstein, 2010; Kirby *et al.*, 2012). Section 3.3 describes a sexual network.

According to Heward et al. (Cooper *et al.*, 2012), rate and frequency are used interchangeably in behavioural studies. For the purpose of this study, we use rate instead of frequency. Stigum et al. (Stigum *et al.*, 1997) found that Norwegian males had higher PTOR compared to females. They used the data from a 1987 survey which was conducted on individuals aged 18 to 60 in Norway and also repeated in 1992 (in Norway) on a new sample of individuals with the same age range as the first. The survey was repeated mainly to study the changes in sexual behaviour since the first survey. Their results suggested that people who have high PTOR are more likely to test for HIV previously. The study also found that recent PTOR (in terms of the number of new sexual partners three years ago) declines with an increase in age. Partner turnover rate was found to be lower in 1992 than in 1987, because the sample subjects taken in the first survey were not the same as those in the second survey, they should make a suggestion for systematic error. They predicted the recent PTOR from earlier PTOR (total number of sexual partners from age 16 to the age three years prior to the survey) and found that the predicted recent number of sexual partners did not depend on age and had no association whatsoever with HIV testing. The males still had a higher PTOR than females. Stigum et al. did not give information about the HIV status of the respondents so we cannot say if the whole sample were HIV positive or not.

Tanfer et al. found that women with high PTOR are more likely to acquire sexually transmitted diseases than women with low PTOR (Tanfer and Schoorl, 1992) and that the risk of HIV infection increases with the number of sexual partners. However, the mode of data collection could cause a bias in their result because the data was collected through personal interviews. Participants may give inaccurate information because they do not trust the interview process to be confidential. Also, women who claimed to be in

long term relationships were not asked if they had other sexual partners, which cannot be ruled out. This could introduce bias into the results. However, the result cannot be generalized since the data only includes the never – married women and so we cannot relate the conclusions to the behaviour of ever-married women of the same cohort (Tanfer and Schoorl, 1992).

May and Anderson (May and Anderson, 1987) also argued that the transmission of HIV depends on PTOR (partner turnover rate). They went on to develop mathematical models, which shows the relationship between this variable and other variables in different populations. They emphasized that the average rate of acquiring new sexual partners is needed to make predictions for HIV infection. These mathematical models are simple models and may not capture all of the details and realities surrounding the transmission dynamics of HIV.

Winkelstein et al. (Winkelstein *et al.*, 1987) used a data set including homosexual and bisexual populations to analyse the association between the reported number of sexual partners (two years before the survey) and HIV serologic status. A strong relationship was detected between number of male sexual partners and seropositivity. However, we cannot be certain of these findings as the response rate of the sample was low (59%).

In their analysis, Mishra et al. (Mishra *et al.*, 2009) analysed data from four demographic health surveys in Zimbabwe, Uganda, Rwanda and Cameroon conducted in 2004-2006. Their results found the evidence of a relationship between the number of lifetime sexual partners and HIV infection in all the countries. The prevalence of HIV was found to increase with the number of lifetime partners and men reported more lifetime sexual partners than women. A strength of this study is that it did not rely on self reported data for HIV status, because HIV tests were carried out on the subjects in the study. Therefore, high partner turnover rate is believed to have great impact on HIV in the society (Temple-Smith, 2014).

3.2.1 Impacts of high PTOR in the society

It fuels an epidemic in the population especially if the core group is large. The core groups are the groups with high PTOR and multiple sexual partners (Borgdorff *et al.*, 1994). It makes the epidemic difficult to control or even uncontrollable (Hamilton and Howard, 2012).

3.2.2 Factors that can reduce PTOR

Delay in sexual debut – early abstinence should be encouraged and publicized through various means of communication. e.g. awareness through sensitization programs, radio and television programs, sexual health messages etc. Shisana *et al.* (2012).
Reduction in the number of sexual partners – this involves reduction in multiple number of partners and concurrency Shisana *et al.* (2012).
Mutual monogamy also contributes to the reduction in high PTOR. This can be achieved within highly regulated societies with enforced social norms, or where sexual partners are trustworthy and faithful to each other (Watts and May, 1992).
Identification of target core groups – If the core groups are quickly identified and intervention policies are implemented, the HIV burden will be reduced Shisana *et al.* (2012).
Interventions targeted at specific sexual networking sites – those increasingly used to find sexual partners, whether physical locations or internet sites – are likely to be most efficient (Temple-Smith, 2014).

3.2.3 PTOR in the global context

In their study, (Wellings *et al.*, 2006) revealed that women in the UK tend to have sexual relationships with twice as many men as they did 20 years ago. It was discovered that women on average now have eight partners in a lifetime and it used to be four sexual partners on average in the 90s. This shows that women in the UK are catching up with men whom on average have 12 sexual partners in a lifetime.
The article reported that adults now have less frequent sex than they used to have about ten years ago. This could be due to more busy schedules in their working lives. In the 90s, both men and women had sex five times a month but has been reduced to three times a month in recent years. This may be due to the fact that fewer people live with their partners (Wellings *et al.*, 2006).
In the US, during the 90s, the National Health and Social Life survey (NHSLS) conducted a study (Laumann, 1994) where over 3,000 persons aged 18 to 59 were sampled (see Table 3.1). They were asked the total number of sexual partners in a lifetime, and the findings revealed that most adults in the US were sexually active. Out of the individuals sampled, 97% of men and women reported having had at least one sexual partner in their lifetime. Many of the male participants reported having ten or fewer sexual partners and of the female participants, 70% reported having four or fewer partners.
Between 2006 and 2010 (Lepkowski *et al.*, 2010), the National Survey of Family Growth (NSFG) carried out a survey that sampled over 13,000 individuals aged between 15 and

44 in the US (see Table 3.2). This finding also revealed that most of the US adults are sexually active and it shows a similar pattern with the results obtained from the 90s as shown in Figures 2.1 and 2.2. Here, more than 90% of men and women were sexually active and 57.3% of the men reported fewer than six lifetime total partners as well as 74.6% of the women (Lehmiller J., 2015).

Table 3.1: Lifetime number of sexual partners in the 90s by the National Health and Social Life Survey (NHSLS)

Number of partners	Percentage of Men Reporting This Number	Percentage of Women Reporting this Number
0	3%	3%
1	20%	31%
2-4	21%	36%
5-10	23%	20%
11-20	16%	6%
21 or more	17%	3%

Table 3.2: Lifetime number of sexual partners in 2006 by the National Survey of Family Growth (NSFG)

Number of partners	Percentage of Men Reporting this Number	Percentage of Women Reporting this Number
0	9.6%	8.6%
1	12.5%	22.5%
2	8.0%	10.8%
3-6	27.2%	32.6%
7-14	19.5%	16.3%
15 or more	23.2%	9.2%

Both studies are similar in the sense that they both revealed that women were more likely to report having only one partner in a lifetime while the men had ten or more. It could be the case that men tend to over-report compared to women who under-report their number of sexual partners due to social desirability bias. Another possibility could be that both sexes view or define sexual partners differently (Lehmiller J., 2015). Note that the ages considered differed across both studies, and the same yardstick may not be used to obtain information.

In 2006/2007, the maker of some of the world's best-selling condoms commissioned a survey to find how customers' sexual behaviour differed between countries and regions. The study was carried out by a primary method called 'Sexual Wellbeing survey'. This

survey method makes it easy for information to be gathered from a large number of people located in different regions. About 26,000 participants aged 16 or older were sampled from 26 countries. Some of the countries are India, the USA, Japan, South Africa, France, Russia, Greece, Brazil, Singapore, Mexico, Germany, Italy, Switzerland, Nigeria, the Netherlands and China. The study revealed that men have more sexual partners than women in all of these countries except Mexico, Austria and some others. Note that these are the average lifetime number of sexual partners. Mexican men had few lifetime number of sexual partners, six on average, and women had an average of 14 sexual partners. Across all the countries surveyed, the average number of sexual partners was 15.9 for men and eight for women. Brazilian men were reported to have had the highest average lifetime number of sexual partners (27) out of all the countries and their women had 11 on average. Austrian men had 17 sexual partners on average while their women had 29. On average, Russians had about 17 to 28 sexual partners and Greeks had about ten to 28 sexual partners, however women had more sexual partners than the men (Plunkettfor, 2014; Tsatsou, 2012). In a study conducted in 2005, average numbers of sexual partners in some selected countries worldwide are shown in Figure 2.3. The next section gives an insight into partner turnover rate in South Africa.

3.2.4 PTOR in SA

It has been reported in the 2012 Durex Global survey that South Africans have 18 lifetime sexual partners on average, just behind Italy with 19, and Spain with 21 (Tsatsou, 2012). In 2005, the Nelson Mandela Foundation commissioned a survey tagged 'South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey' which contained 15, 851 individuals who agreed to HIV testing and were unidentifiably linked to behavioural interviews.

The age group for this survey is 15 years and older. This survey collected data on HIV status of individuals, socio-demographic and behavioural factors that greatly enhanced the analysis and interpretation of the observed trends in HIV prevalence and incidence. Most sexually agile participants reported having one partner during the year with a higher proportion of females (97.4%) reporting this more than the male (83.7%). About 16.3% of males reported having had more sexual partners during the past year and 2.6% females reported the same.

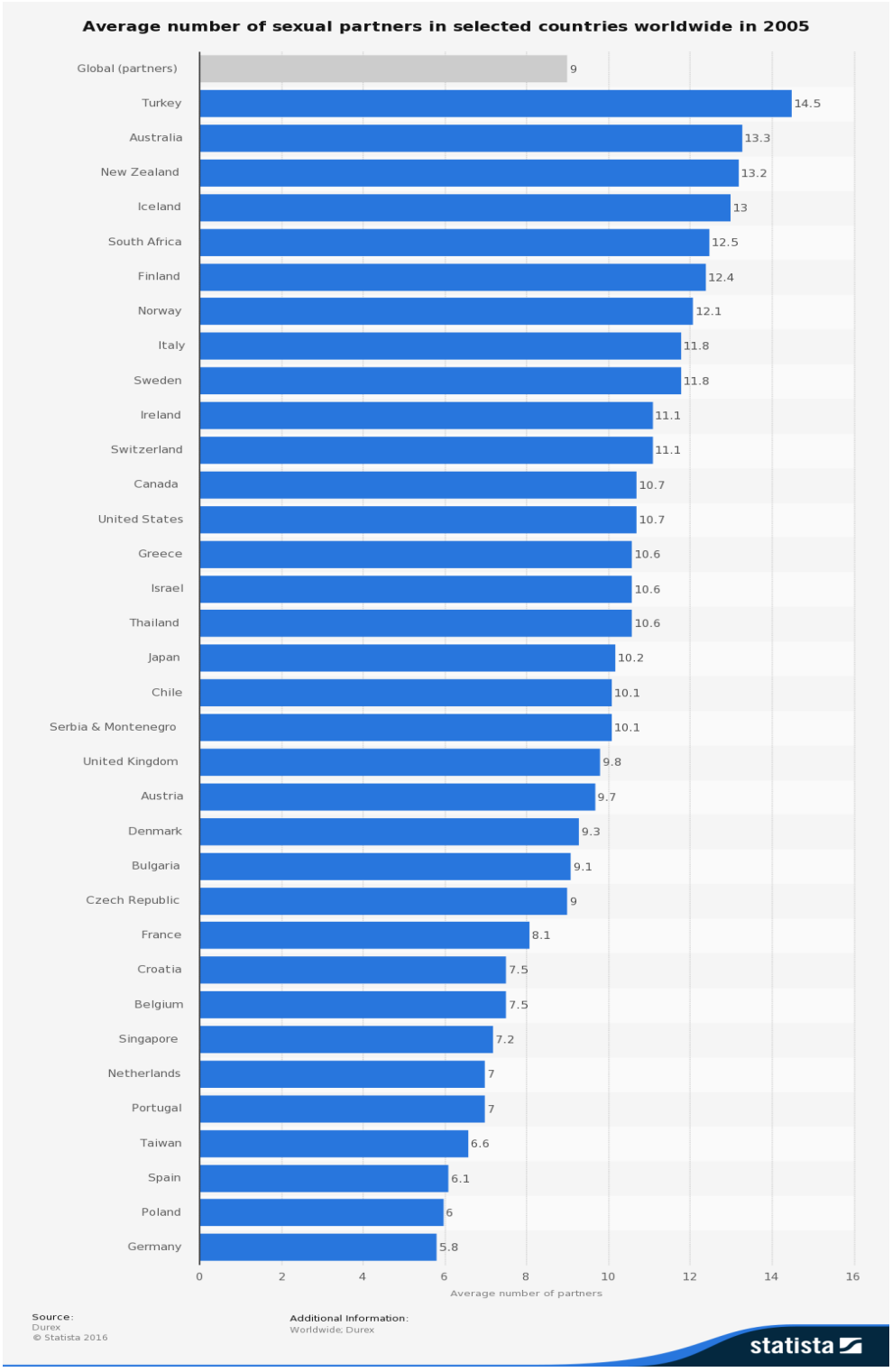


Figure 3.1: Figure showing the result of a survey on the average number of sexual partners in some selected countries in 2005 (Tsatsou, 2012)

Considering the number of sexual partners and age, young males reported a higher number of sexual partners than the older ones, and the same was reported for the females as well, just that the rates were much lower among the females than males. With respect to settlements, the males living in informal areas (both rural and urban) reported more sexual partners than the ones living in formal areas while in the urban areas, the rates reported by the males were higher than their female counterparts living in urban areas. Higher rates of sexual partners were recorded among the African and coloured males while the rates were almost the same in African, coloured and white females but lower for Indian females (Shisana *et al.*, 2005).
Partner turnover rate (PTOR) has been long recognized to play a major part in the transmitting HIV because an infection is known to be likely to persist with high PTOR. Sexual partners are connected to one another and this forms a network.

3.3 Sexual networks

A sexual network consists of group of individuals (nodes) who are sexually connected (edges) to each other. It is a graphical representation of sexual connections of a particular social group. Sexual network study is important for understanding the transmission dynamics of diseases in a population. This type of network is characterized by the number of partners (degree distribution) which indicates the number of ties to others in a network. It shows randomness of partners – diseases spread easily when sexual partners are chosen randomly in a population, the core groups – these are members with highly risky sexual behaviour in a network and centrality, which means an HIV infected person could be the center of a network which fuels the epidemic (Center for AIDS Prevention Studies 2003). A sexual network can be dense: individuals with higher average number of partners or sparse; individuals with lower average number of partners. Analysing sexual networks could help identify the disease transmitters and help explain the causal agents of the epidemic. The measurement variables in a sexual network survey are, the start and end dates of a sexual relationship, encountered location of partners, sexual involvement in relationship, frequency of intercourse, partner concurrency - partners having other sexual partners outside of the ongoing sexual relationship, condom use during copulation, and often other variables as well.
In a sexual network, when one partner is infected with STI, the other partner becomes susceptible and may likely get infected if caution is not taken by using protective methods. This goes on and on in a chain of sexual linkages among individuals. People who are in a network often have no idea of the other persons involved in the network as they

may not be aware of their partners' partners. The different types of patterns of sexual network and behaviour determines how rapidly HIV spreads.
In September 2009, a campaign tagged "The Sexual Network Does Not Stop with you. Get off the Sexual Network!" was launched in Uganda and it addressed the increase in new cases of HIV/AIDS among married couples in Uganda. The objective was to increase serial monogamy among the married, people in long term relationships, well educated people and people living in urban areas. This was aimed at increasing the serial monogamy by five percent from September 2009 to May 2010. The campaign raised awareness about the risks that are associated with multiple and concurrent partnerships. The campaign consisted of three phases. The first phase introduced the concept of sexual network; the second phase explained and educated the individual, individual's family and partner about the consequences of sexual network and the final phase highlighted the steps needed for people to remove themselves from the network. The campaign was propagated through the TV and radio sports programmes, local theatre and call-in radio shows. Presently, the impact data of the campaign has not been made available but the campaign has taken root in social media forums. There exist different forms of sexual partnerships and they are briefly described in the section that follows.

3.4 Sexual partnerships

This type of partnership can be established between same or different sexes, two or more individuals and people of different social status. A relationship may be once-off, which is known as one-night-stand; seasonal, such as extra-marital affair/youthful exuberance; and long – time committed relationship such a nuptials. A sexual relationship can also be exclusive or an open one where the partner is free to engage with multiple sexual partners.
In a sexual relationship, partners may or may not have equal power during a sexual activity. The power in sexual partnership is a function of the social status and personality of the individual forming it. In studies carried out in 1980 and mid 1990s, 75% of sexually active men and women had only one sexual partner in the preceding 12 months (Finer *et al.*, 1999).

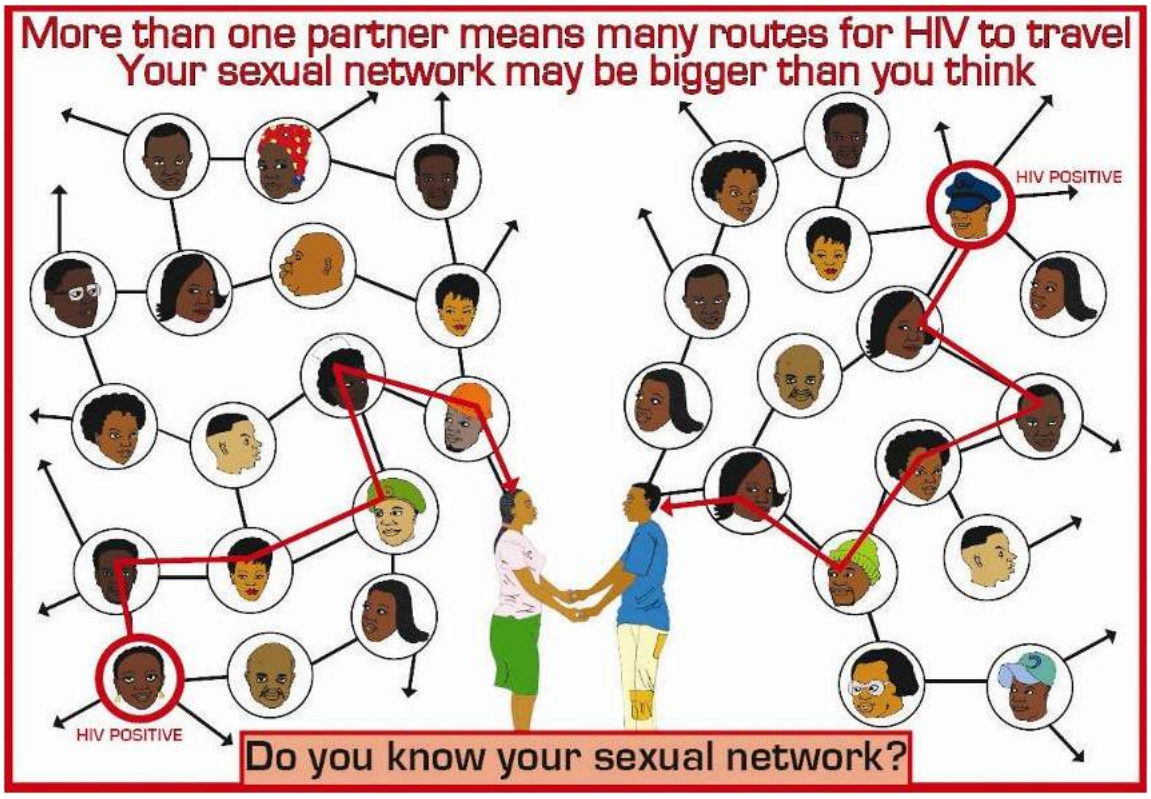


Figure 3.2: Diagram showing a sexual network and how HIV can travel so fast in a network (Delva *et al.*, 2013)

Courtship entails the familiarizing of two individuals. Courtship can be sexual or asexual in nature.

3.5 Reasons for forming new partnerships

Financial motivation and economic conditions can encourage men and women to form newer relationships. Individuals at times want to improve their income or standard of living by bonding with wealthy individuals. In the corporate world, individuals use sexual incentives to climb the corporate ladder and achieve their heart's desires. In other cases, women from poorer parts of the city/slumps indulge in prostitution so they can live above their means (Cohen, 2000).

Sexual satisfaction is a very important factor in a relationship. Anything short of being satisfied sexually may lead to a partner indulging in having concubines or going to sex-workers (Parker *et al.*, 2007).

Prevailing customs and beliefs of a particular society can encourage the formation of

new relationships. In East Africa (Odimegwu and Somefun, 2017), a tribe called Taureg allows women to keep as many husbands as they desire without any backlash.
Bragging rights and social temptations can lead to an individual having multiple sexual partners. Men tend to be very proud of sharing their sexual prowess among colleagues. A man who has a variety of women from different races, sizes and ages tends to have a bloated ego (Meyer-Weitz *et al.*, 2003; Otutubikey Izugbara and Nwabuawele Modo, 2007).
Religious beliefs of a society naturally affect behavioural patterns and choices. Some religions encourage polygamy, making it easier for individuals, especially men, to have as many wives as they can cater for (Thobejane and Flora, 2014).
Intimate partner violence, which also means abusive relationship, can cause an individual to look outside their current relationship to finding new ones (Bell and Naugle, 2005).
Untamed lustful desires - when an individual lusts after another, they tend to abandon their relationship to go for the persons they are lusting after (Fisher, 2000).
Peer pressure - Friends or members of a peer group can influence people to form multiple sexual relationships (Mlambo *et al.*, 2016). Separation/Divorce - A broken home can lead to unfaithfulness among former lovers. Individuals in this type of situation will want to move on if their relationship is irreconcilable, hence they form a newer relationship (Fletcher *et al.*, 2012).
Disability/Illness/Death of one partner can cause the other to seek sexual pleasure elsewhere (Fletcher *et al.*, 2012).

3.5.1 Consequences of new sexual partnerships

The formation of a new relationship increases an individual's risk of contacting sexually transmitted diseases or can help an already infected individual spread diseases among the general populace (Sathiyasusuman *et al.*, 2015). Some individuals form new relationships because of financial situations, but if care is not taken, the more sexual relationships the more the number of offspring. Having children with multiple partners will strain the financial resources of people indulging in it thereby causing artificial or self induced poverty (Cohen, 2000). Few lucky individuals can improve their standard of living by forming multiple sexual relationships. Women from poor neighbourhoods indulging in high class prostitution can live above their means (Cohen, 2000).
Having multiple partners is frowned on in some societies, and indulging in such can bring about divorce and broken marriages (Fletcher *et al.*, 2012).
Religious sentiment labels people indulging in such as outcasts and they are castigated

in the society. This impacts the self esteem of the individual one way or the other (Lamphier and Welch, 2017). Children brought up in an environment where parents indulge in having multiple partners are naturally affected. They may not have a proper home and growing up in such conditions can cause them to view having multiple sexual partners as the norm (Fletcher *et al.*, 2012).

Chapter 4

Design and methodology

4.1 Introduction

The main purpose of the study was to check for inconsistencies in the reported lifetime number of sexual partners and the expected lifetime number of sexual partners, and how these are predictive of HIV status.

4.2 Study design

We conducted a secondary analysis of the data that came from the Cape Town Sexual Behaviour study which was funded by VLIR – FWO. This is a collaborative effort funded by the Flemish Interuniversity council (VLIR – acronym for the Flemish words) and the Research Foundation – Flanders (FWO). The Cape Town Sexual Behaviour study is a cross-sectional study conducted to understand the association between HIV status, sexual connectedness, age disparity and other sexual behavioural factors. It is also to discover behavioural factors that predict egocentric and community sexual network structures Delva (2012). They got their participants using a sampling frame from three communities in the Zambia South Africa TB and AIDS Reduction study (ZAMSTAR) by randomly re-sampling the participants. These communities were Khayelitsha, Delft and Wallacedene.
The ZAMSTAR study was a community – randomised trial aimed at reducing the prevalence of tuberculosis (TB) in communities with a high TB and HIV burden by new public health interventions. HIV tests were carried out and counselling for willing participants were provided by the ZAMSTAR study. This allowed the sexual network data from the Cape Town Sexual Behaviour study to be linked to participants HIV test results in the

ZAMSTAR study. Some participants decided not to test for HIV and this caused a small number of the random sample per study site to consist of non-consenting participants to HIV testing.

4.3 Study setting

The Cape Town Sexual Behaviour study from which we derived the data was conducted from July 2011 to February 2012 in three underprivileged communities in Cape Town: Khayelitsha, Delft and Wallacedene. These communities have a high burden of HIV and were chosen because they had a population of residents who were primarily black or colored, the ethnic groups that are most affected by HIV in South Africa Shisana *et al.* (2012).

4.4 Measuring instruments

A cross-sectional sexual behaviour survey was conducted, using a touch screen questionnaire that utilizes an audio – computer – assisted self – interviews (ACASI) application in obtaining sexual history data.
Participants were asked for basic demographic information and then asked if they had a main sexual partner about a year ago and if they still had the relationship ongoing. They were asked to indicate the periods they were in this relationship on a touch screen time line. The concept to use a touch screen time-line was derived from the Relationship History Calendar (RHC) and an Events History Calendar (EHC) which has been previously certified and evaluated on adolescent and adult participants in sub-Saharan Africa and the US Delva *et al.* (2011); Reniers and Watkins (2010); UNAIDS (1998); Kabiru *et al.* (2010); Martyn and Martin (2003). The time-line recorded the start, cessation and time – span of each relationship period which were shown on the touch screen time line for participants to indicate, using different colours for each partner. Additional questions about the frequency of intercourse, period of engagement in sexual activity with a particular partner (measured in quarter month intervals) and frequency of condom use were derived from the time-line.
They were asked for the year of birth of their partner, and whether they thought their partner had other concurrent partners. They were also asked if the partners or the participants themselves used drugs or alcohol at first intercourse. The level of spatial connectedness was gauged by asking how long the participants took to reach their partner's residence and the degree of proximity of their residences. The questions were asked for

each main sexual partner and casual sexual partners. Participants were asked the number of sexual partners they have ever had at the end of the questionnaire Ayles *et al.* (2008). The maximum number of main partners that they could report were five, and 15 for casual partners. Note that these questions were repeated for as many partners as they wanted to report on.

4.5 Participants

The Cape Town Sexual Behaviour study randomly sampled 1857 people from the ZAMSTAR study, of which 1115 people were located by address (60% contact rate). Of the 742 people whose addresses were not located, 511 relocated to unknown places, 34 people were dead and the reason for non contact of 197 people was unknown after three attempts. 87 participants were excluded due to optical or bodily defects. 878 participants responded out of the 1028 eligible participants left (85.4% response rate).
Of the 878 participants that responded, we excluded participants of races other than black (n= 651), had missing data for age and gender (n= 629), were below 15 years and above 40 years (n= 387). Participants aged below 15 were excluded because we did not trust their ages. The reason for the cut-off at 40 years old is three-fold. Firstly, with increasing age, the risk of recall bias increases for number of lifetime partners. Secondly, with increasing age, the risk of selection bias may also increase if larger number of lifetime or recent partners is associated with HIV infection and subsequently with HIV-related mortality. Finally, the synthetic cohort approach is only valid if the age-specific rates remained constant over the age/time window that we consider. We accumulate these rates from age 16 to age 40. That is a 25 year period. We are implicitly assuming that the age-specific partner turnover rates hardly changed over the past 25 years. Please note that this is a strong assumption and the validity of the conclusion depends on it. If we had a cut-off at 65 years old for example, that assumption would become even stronger (rates staying constant over a 40 years time window). We were left with 387 participants for our analysis.

4.6 Methods

This chapter explains the statistical methods used in line with answering our research questions.
In the first stage of our analysis, we did descriptive statistics which consist of histogram plots showing the number of new sexual partners in the last year and the lifetime num-

ber of partners aggregated by age and gender. Note that this was done separately for men and for women.

The number of new sexual partners in the past year is the number of new relationships started in the year (12 months) before the survey which was measured by adding the number of casual partners and main partners. It is believed to represent their recent sexual behaviour.

The lifetime number of sexual partners measured the total number of sexual partners the participants have had since sexual debut till the present time of the study. It was measured by asking the respondents the total number of sexual partners they have ever had. This is believed to capture the changes in sexual behaviour and it also provides an insight into the risks which the individuals have been exposed to.

In the second stage, we did a Poisson regression to model the relationship between the number of new sexual partners in the last year, lifetime number of sexual partners, age and gender. We went further to check for overdispersion in our data by conducting a dispersion test, this test came out positive, hence overdispersion existed in the data. Because the rate at which individuals form sexual partnerships are different, it is a Poisson process which follows a gamma distribution, we did the negative binomial regression, which follows a Poisson-gamma mixture distribution (details in section 2.5.1).

We did a negative binomial regression to model the relationship between number of new sexual partners in the last year, lifetime number of partners, age and gender. The negative binomial regression is explained in section (2.6). We discovered that age had a non-linear effect on the number of new sexual partners in the past year and the lifetime number of sexual partners.

The natural cubic splines (or restricted splines) were used to allow for the non-linear effect of age in our model (see section (2.7.4). The expected lifetime number of sexual partners resulted from using the model to predict the lifetime number of sexual partners from the reported number of new sexual partners in the last year. There is need to check for inconsistencies between the reported and expected values. We used a synthetic cohort approach to check for inconsistencies between the reported and the expected lifetime number of sexual partners. Section (2.8) discusses this in detail. The differences between the reported (self-reported) and expected values (estimated using the synthetic cohort approach) were plotted and we call this bias, which is a function of age stratified by gender. The bias curve was plotted separately for men and women.

Furthermore, we estimated confidence bounds (95%) around each bias curve by using the bootstrap method. One thousand (1000) bootstrap replications were done to obtain the confidence intervals. This is explained explicitly in section (**??**). The ratio of the num-

ber of new sexual partners in the last year and the lifetime number of sexual partners was computed. Any value greater than one is assumed to be false because the number of new sexual partners in the last year cannot be greater than the lifetime number of partners. This is shown in figure 5.8.
We went further to test for multicollinearity among the predictors, just to ensure there is no correlation effect on the estimates. A negligible amount of correlation was detected.
The last stage tested for the predictive power (ability of a model to correctly anticipate unknown data) of the number of new sexual partners in the last year and the lifetime number of partners on HIV status. We used the modified Poisson regression model to see the relationship between these indicators and HIV status after controlling for the effects of age and gender. We used the leave-one-out cross validation method to estimate the prediction error. These methods are described in sections (2.5.4 and 2.10), respectively.
The indicators - number of new sexual partners in the past year and the lifetime number of sexual partners are believed to have predictive power on HIV status i.e many researchers have used these indicators to predict the risk of having HIV. Note that we only use the data for the respondents who do not have missing values for HIV status (n=306). The aim of this stage is to see, judging from the HIV status we have from the data, if the total number of new sexual partners in the last year and the lifetime number of partners have enough predictive power on HIV status.
We continued by building models, using the modified Poisson regression technique (discussed in section 2.5.4) to predict HIV status (y).
Note that the non-linear effect of age was allowed in these models. f is a restricted cubic spline function with 3 knots t_1, t_2 and t_3.
A reference model was built separately for each gender and its represented as:

$$\text{Model 1: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3,$$

where X_{1i} is the age of subject i and $\pi(y_i)$ is subject i's underlying risk of having HIV.
Total number of new sexual partners in the last year (PLY) is incorporated into model 1 and is given as:

$$\text{Model 2: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3 + \beta_4 X_{2i},$$

where X_{2i} is new number of sexual partners in the last year for subject i.
Furthermore, we added the lifetime number of partners (LNP) to model 1 resulting in a model given as:

$$\text{Model 3: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3 + \beta_4 X_{3i},$$

where X_{3i} is lifetime number of sexual partners for subject i.
We needed to see the combined relationship of these indicators with HIV status, hence the next model:

$$\text{Model 4: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3 + \beta_4 X_{2i} + \beta_5 X_{3i}.$$

These models were built separately for men and women. The leave-one-out cross validation (LOOCV) method was used to select the model with the lowest prediction error (section 2.10). The statistical analysis was done using R version 3.3.1.

4.6.1 Negative Binomial regression

When dealing with extra-Poisson variation (overdispersion) in the regression analysis of count data, a method called negative binomial regression was proposed. In count data analysis, Poisson models are widely used Choi *et al.* (2005); Lee *et al.* (2012), but they are known for not accounting for the overdispersion often displayed by count data. A modification was made to the Poisson model to account for overdispersion. Overdispersion occurs when the variance is larger than the mean, but a special feature of the Poisson distribution is that the variance and the mean are equal, given as: Lee *et al.* (2012)

$$Var\,(Y) = E\,(Y) = \mu.$$

Real life data does not often exhibit this characteristic of the variance being equal to the mean, so we need to account for this by adding a multiplicative random effect to represent the unobserved heterogeneity. This is another type of model for count data where the occurrence probability is a mixture of both Poisson and the Gamma distributions. The notation is given as follows,

$$\Pr\,(y_i|\mu,\alpha) = \frac{\Gamma\left(\alpha^{-1} + y_i\right)}{\Gamma\left(\alpha^{-1}\right) y_i!} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i},$$

where $y_i = 0, 1, 2, \ldots$ and $\alpha \geq 0$, y_i gives the observed values for observation i, Γ is the Gamma function and α is the dispersion parameter.
In negative binomial models, the variance is assumed to be the quadratic function of the mean;

$$\text{Var}\,(Y_i|\boldsymbol{x_i}) \neq E\,(Y_i|\boldsymbol{x_i}) \quad \text{and}$$
$$\mu_i = \exp\left(\boldsymbol{x_i'\beta}\right).$$

The mean and the variance are respectively:

$$E\left(Y_i|\boldsymbol{x_i}\right) = \mu_i = \exp\left(\boldsymbol{x_i'\beta}\right),$$

$$\text{Var}\left(Y_i|\boldsymbol{x_i}\right) = \mu_i\left(1 + \alpha\mu_i\right).$$

For the purpose of our study, the total number of new partners in the last year is represented as **PLY** and the lifetime number of sexual partners represented as **LNP**. The negative binomial regression used to model the relationship between PLY, LNP, age and gender are shown below. Note that these models were constructed separately for men and women. Natural cubic splines (see 4.6.2) were used in the models.

4.6.2 Natural cubic splines

The aim of a model is to simplify the situation as much as possible in order to understand the trends shown by the data Suttinee (2002). When age is modelled as a continuous linear variable, the non-linear effect is not permitted to take place, and the relations that are complex are left out. To permit the non-linear effect of age, we used the natural cubic splines. In our case, natural cubic splines were used over polynomials because of their stable nature when they are interpolated. A spline is a piece-wise polynomial and the pieces are defined by a series of **k** knots which is represented as **k** basis functions Rodrıguez (2001) and the pieces join smoothly at the knots (see section 2.7.4).
In this study, we used **k** equals four. These are the type of splines that restricts the end of a line to a straight one and it prevents the center from distorting the ends. It uses cubic terms in the center of the data. It frees up two degrees of freedom in each boundary and this is considered reasonable because we have less information in this boundary regions according to the following relations Hastie *et al.* (2009):

$$f(y_{\mathbf{PLY}}) = \beta_0 + \beta_1 X + \beta_2(X - t_1)^3 + \beta_3(X - t_2)^3 + \beta_4(X - t_3)^3,$$
$$f(y_{\mathbf{LNP}}) = \beta_0 + \beta_1 X + \beta_2(X - t_1)^3 + \beta_3(X - t_2)^3 + \beta_4(X - t_3)^3,$$

where $y_{\mathbf{PLY}}$ is the number of new partners in the last year, f is a spline function with four knots t_1, t_2, t_3 and t_4. β_0 is the intercept and $\beta_1, \ldots, \beta_4$ are the coefficients.

4.6.3 Modified Poisson regression

We used the modified Poisson regression (models shown in 4.6) in this study because we are interested in directly estimating the risk ratio. Usually, the estimates of the risk ratio are preferrable Zou and Donner (2013) compared to the odds ratio estimates. Risk ratios are more intuitive to understand. This regression technique also provides more conservative results compared to the binomial regression that has a problem of convergence and the estimates obtained are robust to omitted covariates.

4.6.4 Cross Validation

We are interested in knowing how strongly our models have been able to predict the data. We want to see how accurately the models are able to predict new data. However, the leave-one-out cross validation (LOOCV) method is the best feasible method to use because of our small sample size, which allows us to afford the computational power and cost. After obtaining the predictive error estimates using the LOOCV, one thousand (1000) bootstrap replications were carried out to compute the confidence intervals around the differences in the prediction error estimates between the models. This was done because we want to be sure that the differences between the prediction error values are not just due to chance.

Chapter 5

Results

5.1 Introduction

This chapter aims to present the research findings from the secondary analysis done to investigate the reported number of new partners in the last year and the lifetime number of sexual partners. These findings reflect the trend in the different age cohorts and genders examined. The first part reports on our data characteristics. The next part reveals the rate at which each age cohort formed new sexual partnerships in the last year preceding the survey and the lifetime number of sexual partners. The following aspect covers the data analysis using Poisson and negative binomial regression and then compares the results from these approaches. The other aspect of this chapter reports on the inconsistencies in the number of new sexual partners formed in the last year and the lifetime number of sexual partners. The final aspect of this chapter checks for the predictive power of the number of new sexual partners in the last year and the lifetime number of sexual partners on HIV status. Note that our analysis only considered the black participants. The demographic characteristics of the population are presented in section (5.2.1).

5.2 Characteristics of respondents

5.2.1 Socio-demographic characteristics

The participants were aged 16 to 40 years. The median age was 29 years. About 29% were men and 71% were women. Stratified by community, 27% lived in Delft, 44% in Wallacedene and 29% lived in Khayelitsha. Of the study participants, 76% had no job and 24% had jobs. About 84% were Xhosa speakers, 15% were English speakers and

about 1% were Afrikaans speakers. By religion, 64% were Christians, 28% were not religious and about 3% had other religions. Education-wise, about 18% had primary or no education at all, 77% had secondary education and about 5% had tertiary education.

Table 5.1: Proportion of men and women by age groups

Age (in years)	**16-20**	**21-25**	**26-30**	**31-35**	**36-40**
Men	0.10	0.30	0.25	0.21	0.14
Women	0.08	0.23	0.32	0.18	0.19

Using a chi-square test of independence, we tested for difference between the proportion of men and women by age groups.

Hypothesis 1a. *There is no difference in the proportion of men and women,* $P_M = P_W$

Hypothesis 1b. *There is a difference in the proportion of men and women,* $P_M \neq P_W$,

where P_M and P_W are the proportions of men and women respectively. We use $\alpha = 0.05$.
We obtained a p-value of $0.2772 > 0.05$ and X-squared value of 5.099. Therefore, we do not reject the null hypothesis and conclude that there is no difference in the proportion of men and women by age groups.

5.3 Sexual behaviour

The total number of new partners in the last year (NPLY) varied from zero to 11. A higher percentage of women had zero new partners in the last year preceding the survey, and a higher percentage of men had more than six new sexual partners in the last year. This sample confirms that women report lower number of sexual partners in the last year compared to men. This analysis was done separately for each gender. Table 2 shows the number of new sexual partners in the last year (NPLY) disaggregated by gender (in proportion).

Using a chi-square test of independence, we tested for difference between the proportion of men and women by gender.

Hypothesis 2a. *There is no difference in the distribution of NPLY and LNP for men and women,* $P_M = P_W$

Hypothesis 2b. *There is a difference in the distribution of NPLY and LNP for men and women,* $P_M \neq P_W$,

Table 5.2: NPLY by gender (in %)

	Men	**Women**
0	48	77
1	15	10
2	14	8
3	10	2
4	5	1
5	3	1
6+	5	1

Table 5.3: LNP by gender (in %)

	Men	**Women**
0	0	0
1	14	13
2	12	21
3	12	23
4	15	11
5	9	10
6+	38	22

where P_M and P_W are the proportions of men and women respectively. We use $\alpha = 0.05$.

We obtained a p-value of $1.00 > 0.05$ (for both) and X-squared 0.235 and 0.165 for both NPLY and LNP by gender respectively. Therefore, we do not reject the null hypothesis and conclude that there is no evidence to support that the distribution of NPLY and LNP is different for men and women.

The lifetime number of partners (LNP) varied from one to 15. Men reported more lifetime number of sexual partners than women. Of the men, none in age group 16 to 20 years old were HIV positive, a higher percentage are HIV positive in age group 36 to 40. As for the women, they got infected earlier with 1% aged 16 to 20 years old HIV positive individuals, a higher percentage (16%) of the HIV infected women are aged 26 to 30.

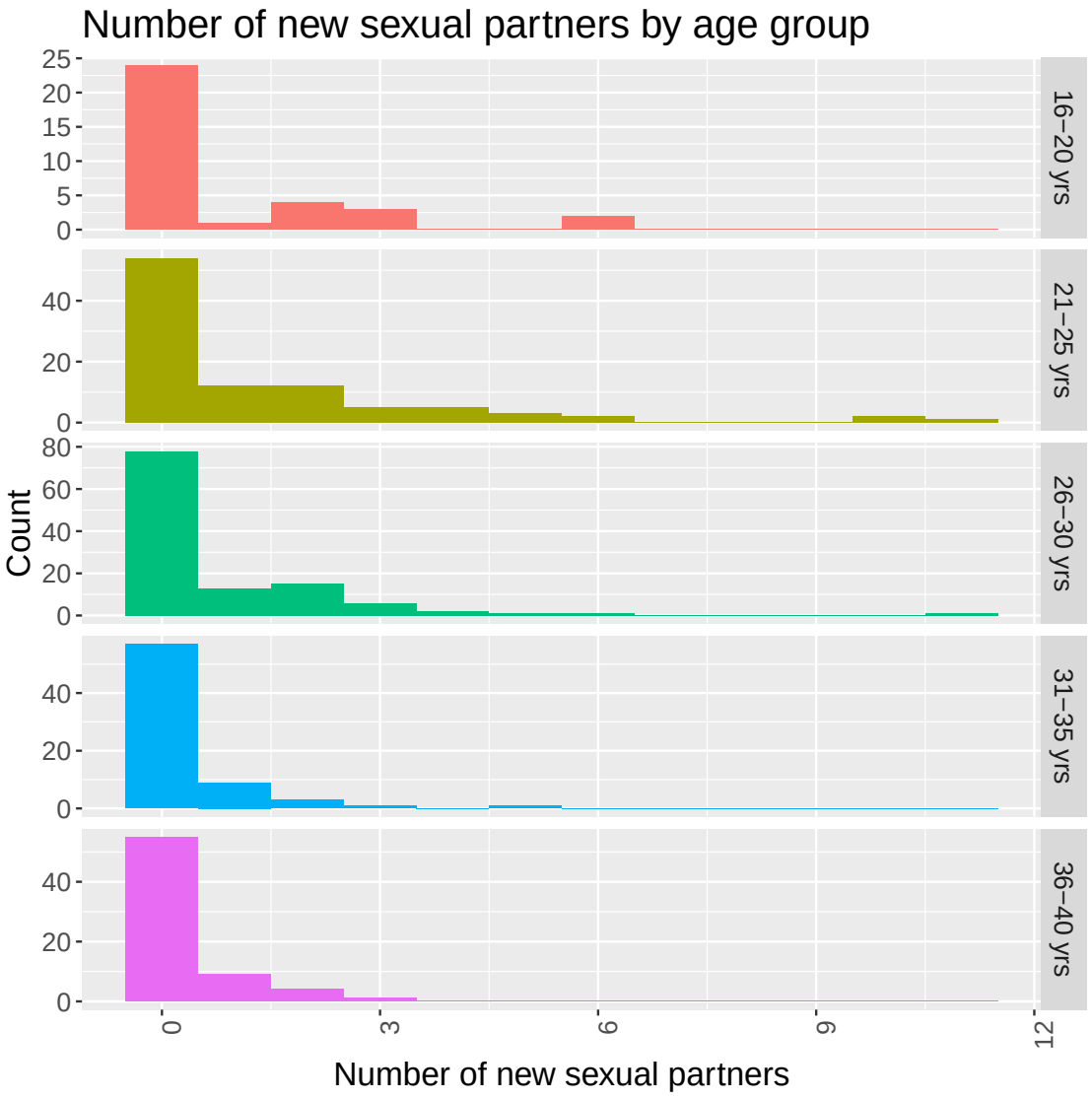


Figure 5.1: Sexual partners by age

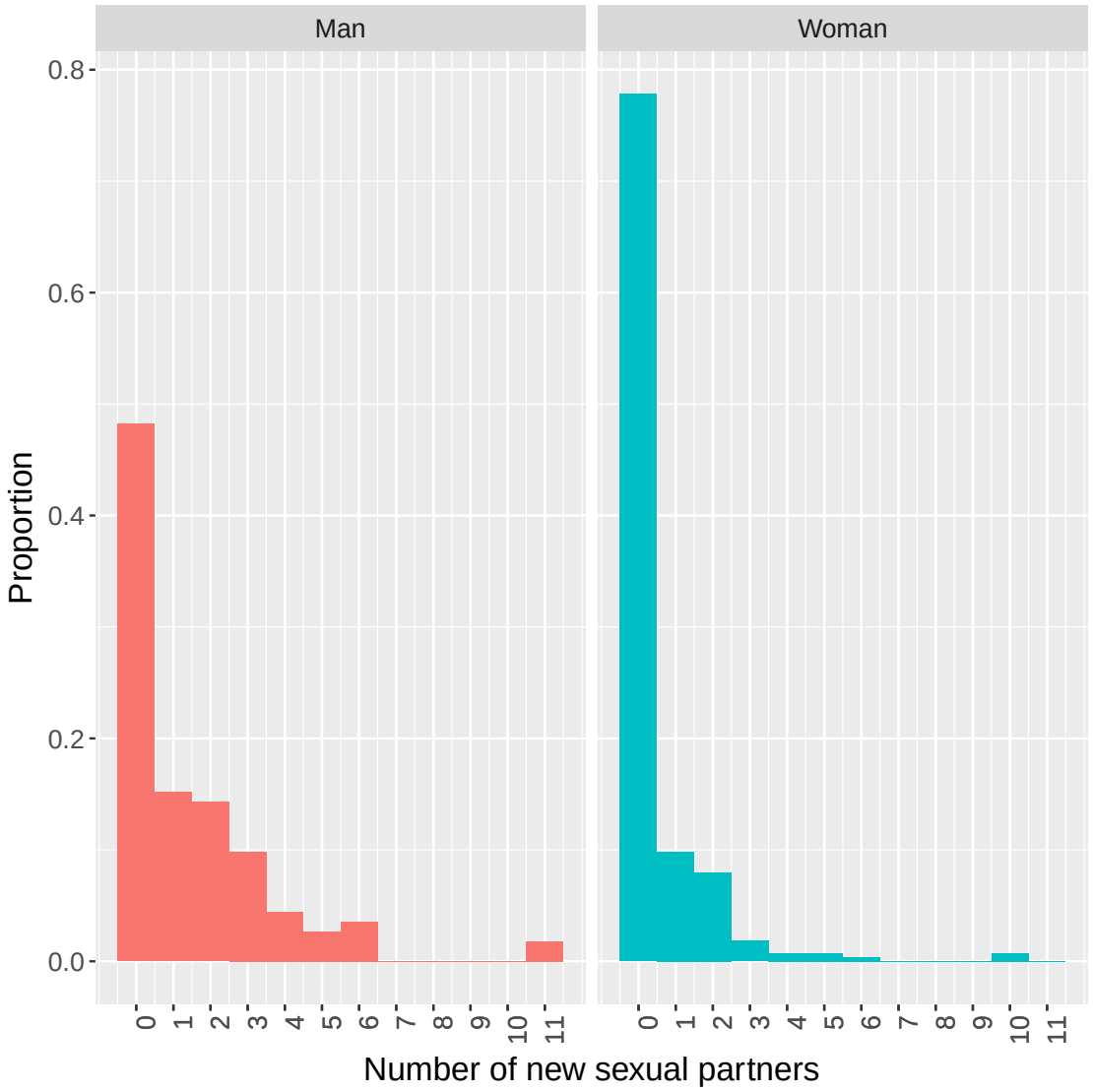


Figure 5.2: Sexual partners by gender

Figure 5.2 shows the total number of new partners in the last year stratified by gender and Figure 5.1 displays the total number of new partners in the last year as stratified by age. It can be observed from Figure 5.2 that most of the women reported zero number of new sexual partners in the last year and Figure 5.1 indicates that most of the individuals in age-group 26 to 30 years had zero number of new sexual partners. Figures 5.4 and 5.3 shown below presents the lifetime number of partners by gender and age group. It is evident from Figure 5.4 that more women reported lesser number of lifetime sexual partners than the men and Figure 5.3 shows that individuals aged 16 to 20 years had a

similar distribution of lifetime number of sexual partners with their older counterparts aged 31 to 35 years.

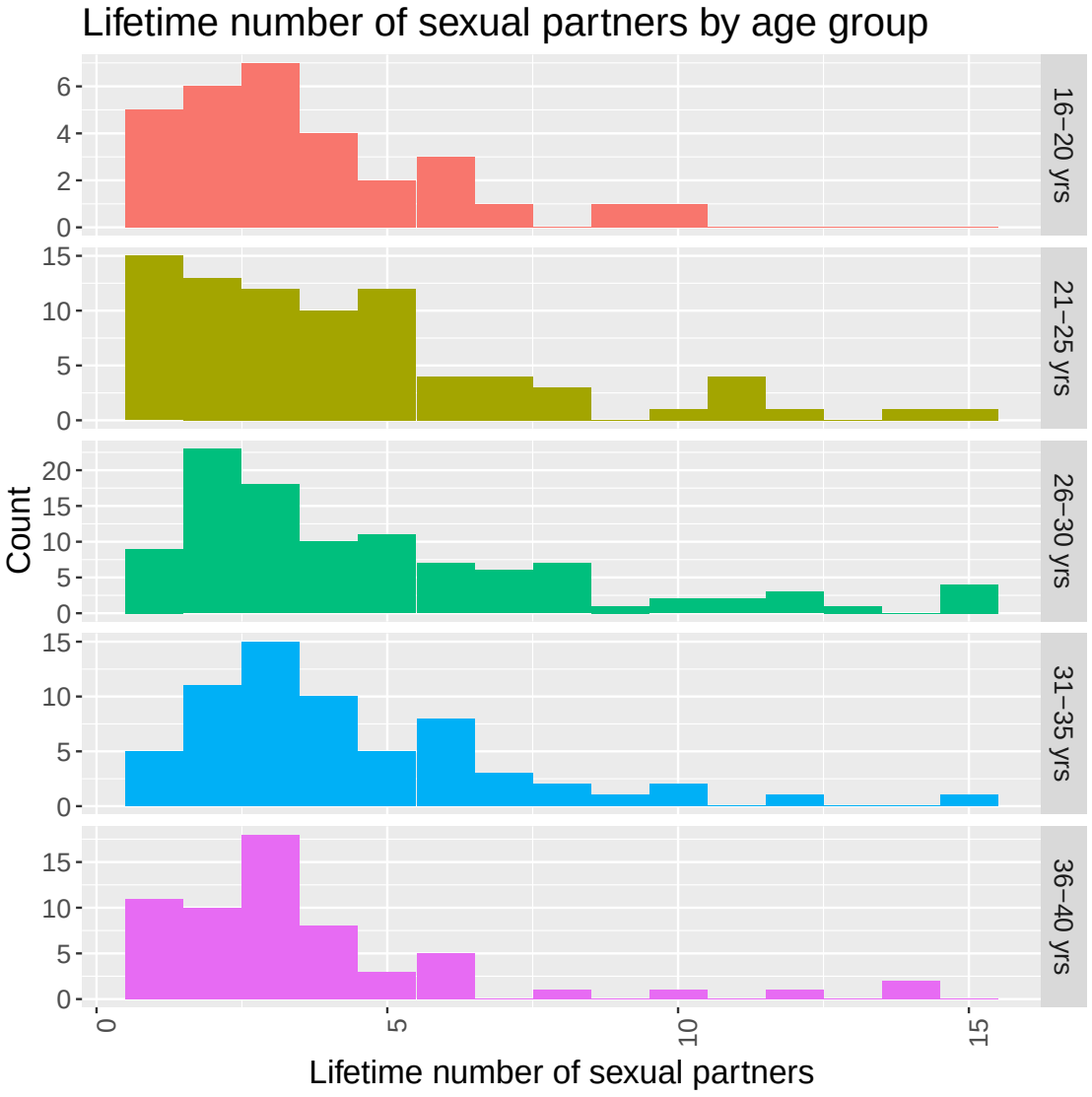


Figure 5.3: Lifetime partners by age

Our analysis shows that the men reported a higher lifetime number of sexual partners than women and the younger cohort aged 21 to 35 on average had a higher lifetime number of partners than their older counterparts aged 36 to 40 years.

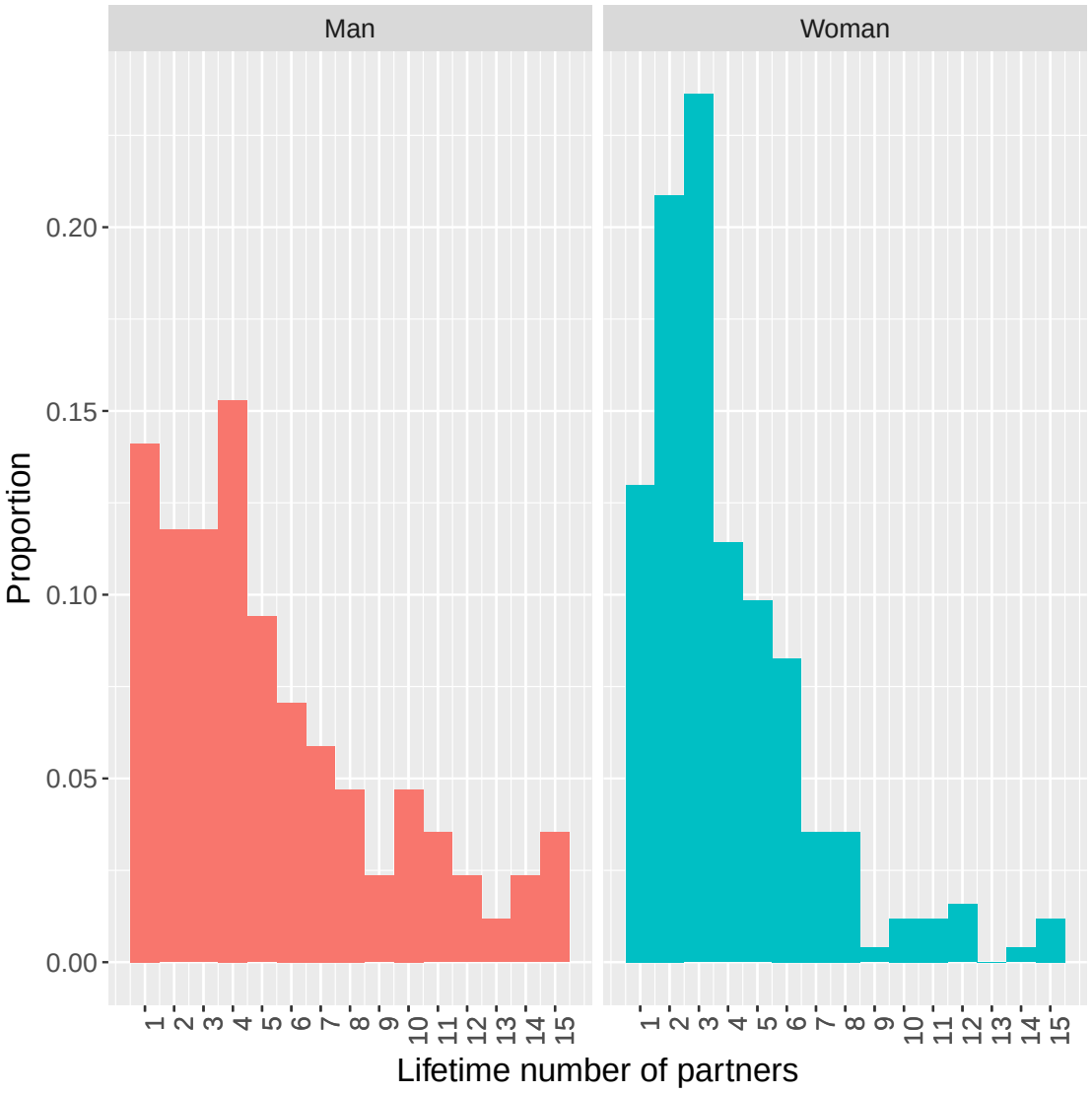


Figure 5.4: Sexual partners by gender

5.4 Model results

5.4.1 Non-linear effect of age

To accurately model the effect of age, the natural cubic spline method was used. This flexibility was allowed due to the fact that the effect of age could be positive up to a certain age, and then becomes negative thereafter. Therefore, instead of assuming a linear effect of age, the effect of differing ages was modelled.

5.4.2 Data analysis using Poisson regression

Due to the fact that our data is a count data, we started off by doing a Poisson regression. Our response variable is the number of new sexual partners in the last year (NPLY – as shown in equation 5.4.1) and the covariate is age. Note that the regression analysis was carried out separately for each gender.

$$\text{Model 1:} \quad f(log(\mu_{NPLY})) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)^3 + \beta_3 (X - t_2)^3 + \beta_4 (X - t_3)^3 \quad (5.4.1)$$

$$\text{Model 2:} \quad f(log(\mu_{LNP})) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)^3 + \beta_3 (X - t_2)^3 + \beta_4 (X - t_3)^3 \quad (5.4.2)$$

Equation 5.4.2 shows the Poisson regression model with lifetime number of partners (LNP) as the response variable. f represents the spline function, μ_{NPLY} is the expected new number of sexual partners in the last year, X is the covariate "age", t_1, t_2 and t_3 are the knots, β_1 is the intercept and $\beta_1 \dots \beta_4$ are the coefficients. Note that we chose the number of knots arbitrarily, which in this case is 4.

Table 5.4: Poisson model estimates from Model 1 for the male population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.6660	0.2390	2.786	0.00533
ns(age, 4)1	-0.1840	0.3537	-0.520	0.6029
ns(age, 4)2	-0.7689	0.4985	-1.543	0.1229
ns(age, 4)3	-1.5390	0.7016	-2.194	0.0283
ns(age, 4)4	-2.2791	0.7710	-2.956	0.0031

Table 5.4 displays the estimates obtained from Model 1 for the male stratum. This means that for a unit increase in age between ages 19 to 23, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(-0.1840) = 0.8319$, between ages 23 to 28, the expected NPLY increases by 0.4635, between ages 28 to 33, the expected NPLY increases by 0.2146 and between ages 33 to 40, the expected NPLY increases by 0.1024. As the number of knots increased, the standard errors increased as well.

Table 5.5: Poisson model estimates from Model 2 for the male population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.3487	0.1797	7.506	$6.1e^{-14}$
ns(age, 4)1	0.3650	0.2277	1.603	0.1088
ns(age, 4)2	0.0378	0.2323	0.162	0.8709
ns(age, 4)3	0.8260	0.4529	1.824	0.0682
ns(age, 4)4	-0.0218	0.2310	-0.094	0.9247

Table 5.5 shows the estimates obtained from Model 2 for the male stratum. This means that for a unit increase in age between ages 19 to 23, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(0.3650) = 1.4405$, between ages 23 to 28, the expected NPLY increases by 1.0385, between ages 28 to 33, the expected NPLY increases by 2.2842 and between ages 33 to 40, the expected NPLY increases by 0.9784.

Table 5.6: Poisson model estimates from Model 1 for the female population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.8557	0.4889	-1.750	0.0801
ns(age, 4)1	-0.0928	0.4609	-0.201	0.8404
ns(age, 4)2	-1.7297	0.6623	-2.612	0.0090
ns(age, 4)3	0.2318	1.2206	0.190	0.8494
ns(age, 4)4	-1.0689	0.4423	-2.416	0.0157

Table 5.6 displays the estimates obtained from Model 1 for the female stratum. This means that for a unit increase in age between ages 17 to 24, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(-0.0928) = 0.9114$, between ages 24 to 29, the expected NPLY increases by 0.1773, between ages 29 to 33, the expected NPLY increases by 1.2609 and between ages 33 to 40, the expected NPLY increases by 0.3434. As the number of knots increases, the standard errors increased and only reduced on the $4th$ knot.

Table 5.7: Poisson model estimates from Model 2 for the female population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.9556	0.2215	4.315	$1.59e^{-5}$
ns(age, 4)1	0.6930	0.2055	3.372	0.0008
ns(age, 4)2	0.1318	0.1932	0.682	0.4951
ns(age, 4)3	0.6123	0.5248	1.167	0.2433
ns(age, 4)4	0.0822	0.1422	0.578	0.5632

Table 5.7 shows the estimates obtained from Model 2 for the female stratum. This means that for a unit increase in age between ages 17 to 24, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(0.6930) = 1.9997$, between ages 24 to 29, the expected NPLY increases by 1.1409, between ages 29 to 33, the expected NPLY increases by 1.8447 and between ages 33 to 40, the expected NPLY increases by 1.0857.

5.4.3 Overdispersion

Equidispersion, which occurs when the conditional variance is equal to the mean is one of the properties of a Poisson distribution. This is unusual in practical data, therefore the data has to be tested.
We went further to test for overdispersion in our Poisson model and we found that overdispersion was present in the data as shown below:

Hypothesis 3a. *There is no overdispersion in the data,* $c = 0$

Hypothesis 3b. *Overdispersion is present in the data,* $c > 0$,

where c is the dispersion term. We use $\alpha = 0.05$

Table 5.8: p-values from the dispersion test

	Model 1	Model 2
Male	0.005	$2.3e^{-6}$
Female	0.002	$2.0e^{-4}$

From Table 5.8, our resulting p-values are < 0.05. Therefore, we reject the null hypothesis (3a) and conclude that overdispersion is present in the data.
This prompted us to consider using a negative binomial model so as to control for overdispersion.

5.4.4 Data analysis using negative binomial regression

It was observed that the data stems from a negative binomial distribution and we stratified the data by gender. The models were built separately for each gender cohort as per the number of new sexual partners in the last year (NPLY) and the lifetime number of sexual partners (LNP) as a function of age (refer to section 4.6). In this study, age was not forced to be linear but a natural cubic spline technique was used to represent the non-linear effect of age (section 4.6.2). Note that $\mathbf{k} = 4$.

$$\text{Model A:} \quad f(ln(\mu_{\mathbf{NPLY}})) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)^3 + \beta_3 (X - t_2)^3 + \beta_4 (X - t_3)^3, \quad (5.4.3)$$

$$\text{Model B:} \quad f(ln(\mu_{\mathbf{LNP}})) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)^3 + \beta_3 (X - t_2)^3 + \beta_4 (X - t_3)^3, \quad (5.4.4)$$

where $\mu_{\mathbf{NPLY}}$ is the expected number of new partners in the last year, $\mu_{\mathbf{LNP}}$ is the expected lifetime number of sexual partners, f is a spline function with four knots t_1, t_2, t_3 and t_4. β_0 is the intercept and $\beta_1, \ldots, \beta_4$ are the coefficients.

Table 5.9: Estimates from Model A for the male population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.6446	0.4363	1.48	0.1395
ns(age, 4)1	-0.2455	0.6052	-0.41	0.6850
ns(age, 4)2	-0.5618	0.7229	-0.78	0.4370
ns(age, 4)3	-1.5685	1.1768	-1.33	0.1826
ns(age, 4)4	-2.5661	1.0101	-2.54	0.0111

Table 5.9 presents the estimates obtained from Model A for the men. This means that for a unit increase in age between ages 19 to 23, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(-0.2455) = 0.7823$, between ages 23 to 28, the expected NPLY increases by 0.5702, between ages 28 to 33, the expected NPLY increases by 0.2084 and between ages 33 to 40, the expected NPLY increases by 0.0768. There was a decrease in the value of the standard error when four knots were used.

Table 5.10: Estimates from Model B for the male population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.3582	0.2711	5.01	0.0000
ns(age, 4)1	0.3601	0.3609	1.00	0.3184
ns(age, 4)2	0.0308	0.3646	0.08	0.9326
ns(age, 4)3	0.8005	0.6961	1.15	0.2502
ns(age, 4)4	-0.0187	0.3703	-0.05	0.9597

Table 5.10 displays the estimates obtained from Model B for the male population in which there was a decrease in the standard error when four knots were used. This means that for a unit increase in age between ages 19 to 23, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(0.3601) = 1.4335$, between ages 23 to 28, the expected NPLY increases by 1.0313, between ages 28 to 33, the expected NPLY increases by 2.2267 and between ages 33 to 40, the expected NPLY increases by 0.9815.

Table 5.11 shows the estimates obtained from Model A for the women. This means that for a unit increase in age between ages 17 to 24, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(0.0771) = 1.0802$, between ages 24 to 29, the expected NPLY increases by 0.1387, between ages 29 to 33, the expected NPLY increases by 1.0140 and between ages 33 to 40, the expected NPLY increases by 0.3867.

Table 5.11: Estimates from Model A for the female population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.8276	0.8669	-0.95	0.3398
ns(age, 4)1	0.0771	0.8203	0.09	0.9251
ns(age, 4)2	-1.9756	0.9525	-2.07	0.0381
ns(age, 4)3	0.0139	2.1283	0.01	0.9948
ns(age, 4)4	-0.9502	0.6435	-1.48	0.1398

Table 5.12: Estimates from Model B for the female population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.9698	0.2690	3.60	0.0003
ns(age, 4)1	0.6876	0.2514	2.73	0.0062
ns(age, 4)2	0.1182	0.2381	0.50	0.6196
ns(age, 4)3	0.5734	0.6427	0.89	0.3723
ns(age, 4)4	0.0851	0.1756	0.48	0.6280

Table 5.12 reveals the estimates obtained from Model B for the women. This means that for a unit increase in age between ages 17 to 24, the expected number of new sexual partners in the last year preceding the survey (NPLY) increases by $\exp(0.6876) = 1.9889$, between ages 24 to 29, the expected NPLY increases by 1.1255, between ages 29 to 33, the expected NPLY increases by 11.7743 and between ages 33 to 40, the expected NPLY increases by 1.0888.

The negative binomial regression fixed the problem of overdispersion shown by the Poisson regression. Next, we consider the relationship between the Poisson and the negative binomial regression.

5.4.5 Poisson versus Negative binomial regression

5.4.5.1 Comparing estimates

We compare the estimates obtained from both Poisson and negative binomial regression in the table below:

Table 5.13: Estimates from Poisson and Negative binomial models for the male stratum

	Poisson	Negative binomial
(Intercept)	0.6660	0.6446
ns(age, 4)1	-0.1840	-0.2455
ns(age, 4)2	-0.7689	-0.5618
ns(age, 4)3	-1.5390	-1.5685
ns(age, 4)4	-2.2791	-2.5661

Table 5.13 shows the estimates obtained from a Poisson and negative binomial regression respectively for the male population, where the new number of sexual partners in the last year is the outcome variable.

Table 5.14: Estimates from Poisson and Negative binomial models for the female stratum

	Poisson	Negative binomial
(Intercept)	-0.8557	-0.8276
ns(age, 4)1	-0.0928	0.0771
ns(age, 4)2	-1.7297	-1.9756
ns(age, 4)3	0.2318	0.0139
ns(age, 4)4	-1.0689	-0.9502

Table 5.14 shows the estimates obtained from a Poisson and negative binomial regression respectively for the female population, where the new number of sexual partners in the last year is the outcome variable.

Table 5.15: Coefficients from Poisson and Negative binomial models for the male stratum

	Poisson	Negative binomial
(Intercept)	1.3487	1.3582
ns(age, 4)1	0.3650	0.3601
ns(age, 4)2	0.0378	0.0308
ns(age, 4)3	0.8260	0.8005
ns(age, 4)4	-0.0218	-0.0187

Table 5.15 shows the estimates obtained from a Poisson and negative binomial regression respectively for the male population, where the lifetime number of sexual partners is the outcome variable.

Table 5.16: Coefficients from Poisson and Negative binomial models for the female stratum

	Poisson	Negative binomial
(Intercept)	0.9556	0.9698
ns(age, 4)1	0.6930	0.6876
ns(age, 4)2	0.1318	0.1182
ns(age, 4)3	0.6123	0.5734
ns(age, 4)4	0.0822	0.0851

Table 5.16 shows the estimates obtained from a Poisson and negative binomial regression respectively for the female population, where the lifetime number of sexual partners is the outcome variable.

Summary

The estimates from both Poisson and negative binomial regression are no much different from each other. This means that they both give consistent estimates of parameter coefficients.

5.4.5.2 Comparing standard error estimates

We compare the standard error estimates obtained from both Poisson and negative binomial regression in the table below:

Table 5.17: Standard error estimates from Poisson and Negative binomial models for the male stratum

	Poisson	Negative binomial
(Intercept)	0.2390	0.4363
ns(age, 4)1	0.3537	0.6052
ns(age, 4)2	0.4985	0.7229
ns(age, 4)3	0.7016	1.1768
ns(age, 4)4	0.7710	1.0101

Table 5.17 shows the standard error estimates obtained from a Poisson and negative binomial regression respectively for the male population, where the new number of sexual partners in the last year is the outcome variable.

Table 5.18: Standard error estimates from Poisson and Negative binomial models for the female stratum

	Poisson	Negative binomial
(Intercept)	0.4889	0.8669
ns(age, 4)1	0.4609	0.8203
ns(age, 4)2	0.6623	0.9525
ns(age, 4)3	1.2206	2.1283
ns(age, 4)4	0.4423	0.6435

Table 5.18 shows the standard error estimates obtained from a Poisson and negative binomial regression respectively for the female population, where the new number of sexual partners in the last year is the outcome variable.

Table 5.19: Standard error estimates from Poisson and Negative binomial models for the male stratum

	Poisson	Negative binomial
(Intercept)	0.1797	0.2711
ns(age, 4)1	0.2277	0.3609
ns(age, 4)2	0.2323	0.3646
ns(age, 4)3	0.4529	0.6961
ns(age, 4)4	0.2310	-0.3703

Table 5.19 shows the standard error estimates obtained from a Poisson and negative binomial regression respectively for the male population, where the lifetime number of sexual partners is the outcome variable.

Table 5.20: Standard error estimates from Poisson and Negative binomial models for the female stratum

	Poisson	Negative binomial
(Intercept)	0.2215	0.2690
ns(age, 4)1	0.2055	0.2514
ns(age, 4)2	0.1932	0.2381
ns(age, 4)3	0.5248	0.6427
ns(age, 4)4	0.1422	0.1756

Table 5.20 shows the standard error estimates obtained from a Poisson and negative binomial regression respectively for the female population, where the lifetime number of sexual partners is the outcome variable.

Summary

The standard errors generated by the Poisson regression is smaller than those generated by the negative binomial regression in all the four models. This agrees with the existing fact that errors are underestimated by Poisson regression.

5.4.5.3 Comparing confidence intervals

We compare the confidence intervals obtained from both Poisson and negative binomial regression in the table below:

Table 5.21 compares the 95% confidence intervals for both Poisson and negative binomial regression. This is when our outcome variable is the new number of sexual partners in the last year with age as covariate for the male stratum.

Table 5.21: Confidence interval estimates for Poisson and Negative binomial model for the male population

	Poisson		Negative	binomial
	2.5%	97.5%	2.5%	97.5%
(Intercept)	0.1688	1.1084	-0.2251	1.5924
ns(age, 4)1	-0.8851	0.5048	-1.4333	0.9304
ns(age, 4)2	-1.7583	0.2065	-2.0767	1.0354
ns(age, 4)3	-2.9199	-0.1489	-4.0527	0.7797
ns(age, 4)4	-4.0392	-0.9709	-4.9962	-0.6391

Table 5.22: Confidence interval estimates for Poisson and Negative binomial model for the female population

	Poisson		Negative	binomial
	2.5%	97.5%	2.5%	97.5%
(Intercept)	-1.8825	0.0391	-2.6688	1.1569
ns(age, 4)1	-0.9690	0.8469	-1.7522	1.8028
ns(age, 4)2	-3.0914	-0.4773	-4.0250	-0.0396
ns(age, 4)3	-2.0541	2.7433	-4.7262	4.5209
ns(age, 4)4	-2.0142	-0.2626	-2.2334	0.3174

Table 5.22 juxtaposes the 95% confidence intervals for both Poisson and negative binomial regression. This is when our outcome variable is the new number of sexual partners in the last year with age as covariate for the female stratum.

Table 5.23: Confidence interval estimates for Poisson and Negative binomial model for the male population (lifetime partners)

	Poisson		Negative	binomial
	2.5%	97.5%	2.5%	97.5%
(Intercept)	0.9814	1.6866	0.8337	1.8890
ns(age, 4)1	-0.0805	0.8131	-0.3461	1.0697
ns(age, 4)2	-0.4198	0.4919	-0.6863	0.7517
ns(age, 4)3	-0.0386	1.7390	-0.5524	2.1469
ns(age, 4)4	-0.4942	0.4135	-0.7471	0.7293

Table 5.23 compares the 95% confidence intervals for both Poisson and negative binomial regression. This is when our outcome variable is the lifetime number of sexual partners with age as covariate for the male stratum.

Table 5.24: Confidence interval estimates for Poisson and Negative binomial model for the female population (lifetime partners)

	Poisson		Negative	binomial
	2.5%	97.5%	2.5%	97.5%
(Intercept)	0.5066	1.3752	0.4403	1.4817
ns(age, 4)1	0.3008	1.1073	0.2048	1.1830
ns(age, 4)2	-0.2446	0.5132	-0.3457	0.5855
ns(age, 4)3	-0.3909	1.6676	-0.6537	1.8326
ns(age, 4)4	-0.1999	0.3581	-0.2608	0.4281

Table 5.24 compares the 95% confidence intervals for both Poisson and negative binomial regression. This is when our outcome variable is the lifetime number of sexual partners with age as covariate for the female stratum.

Summary

The confidence interval estimates produced by the negative binomial regression are wider compared to the ones produced by Poisson regression. This shows that even though the negative binomial regression gives consistent parameter estimates and standard error values, the wider confidence interval estimates indicate less precision in the resulting estimates.

5.5 Inconsistencies in reported and expected values

For the purpose of this study, a synthetic cohort approach was used to estimate the expected lifetime number of partners based on the reported number of new sexual partners in the last year.

The predicted values (from the model) of the lifetime number of partners were used to compare the reported values of the lifetime number of sexual partners (section **??**).

Figure (5.5a) shows the reported number of new sexual partners in the last year for both men and women, and Figure (5.5b) shows the reported lifetime number of sexual partners by age for both gender (from the data). These curves represent model predictions, it is the expected number of partners for a given value of age and gender. A decline was observed in the reported number of new sexual partners for men at age 23 and for women at age 24. The younger cohort formed a higher number of sexual partners than the older cohort.

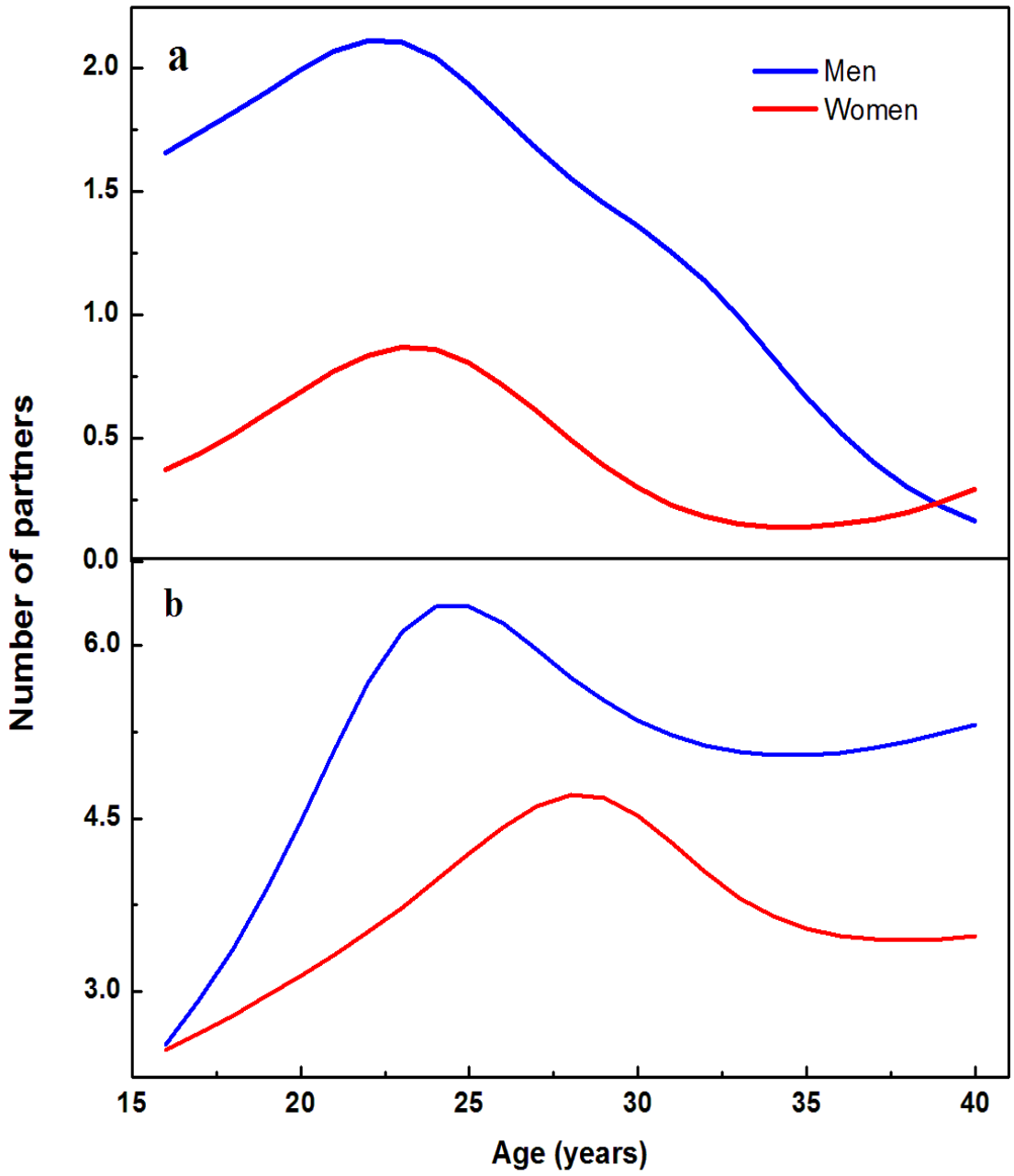


Figure 5.5: (a) shows the reported number of new sexual partners in the last year for men (blue line) and women (red line) and (b) shows the reported lifetime number of sexual partners for both men (blue line) and women (red line). This is as observed from the data. Please note the difference in the scales along the y-axis.

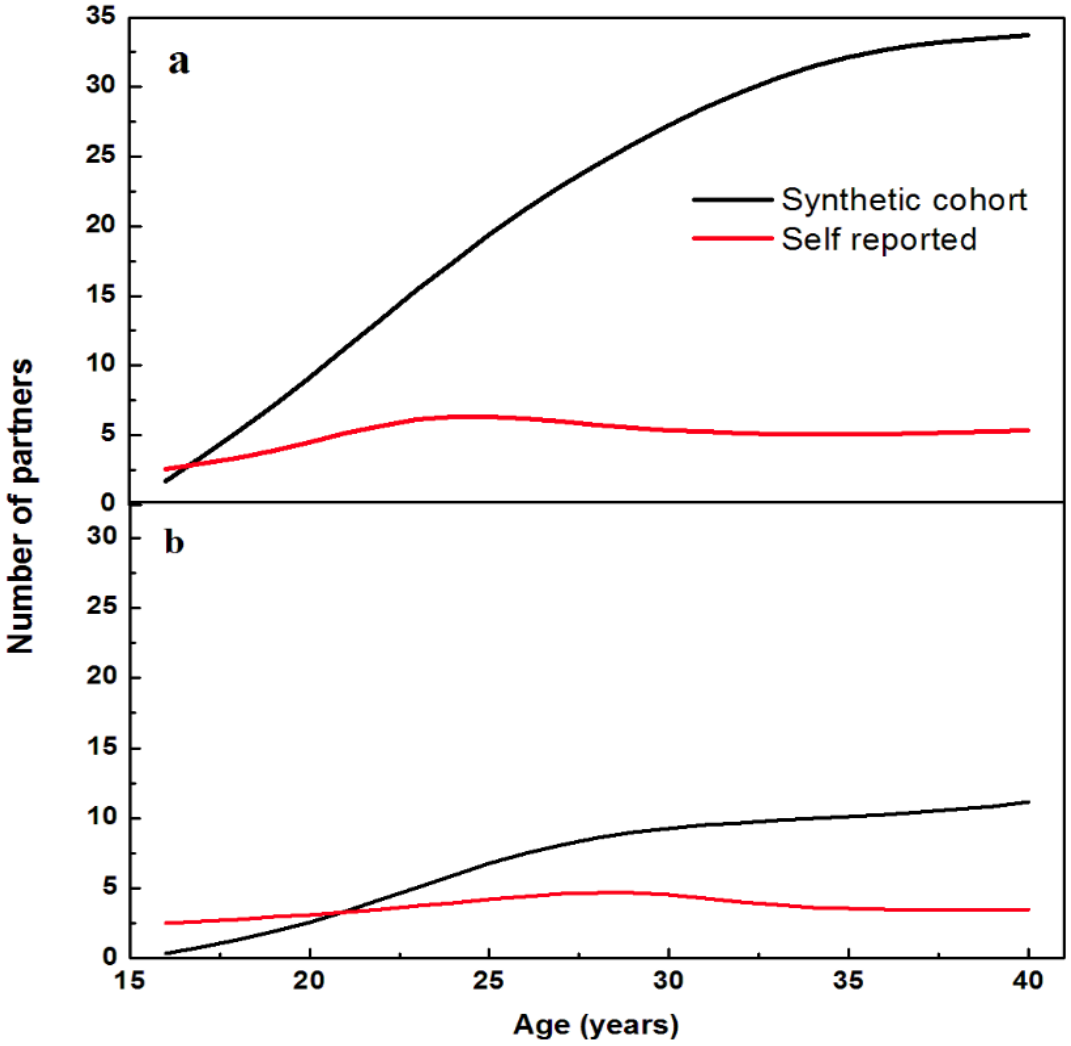


Figure 5.6: (a) shows the reported (red line) and the expected (black line) lifetime number of sexual partners for men and (b) shows the reported (red line) and the expected (black line) lifetime number of partners for women as a function of age, using the synthetic cohort approach. The black line shows the result of the synthetic cohort approach while the red line only shows the reported information as observed from the data.

It is obvious from Figure (5.6a) that the reported and the expected lifetime number of sexual partners for men are not consistent. There exists a huge discrepancy between them. Figure (5.6a) shows that there is a very huge difference in the self reported and the expected lifetime number of sexual partners as age increases. Also, for women in Figure (5.6b), there is a discrepancy in the reported and the expected number of sexual partners as age increases.

5.5.1 Result from the bootstrap method

In order to check if there is enough evidence to see if the discrepancy between the reported and expected number of sexual partners is more than just noise, and to quantify the bias in the data, we calculated the differences between the reported and expected lifetime number of sexual partners for both men and women as shown in Figure 5.7.

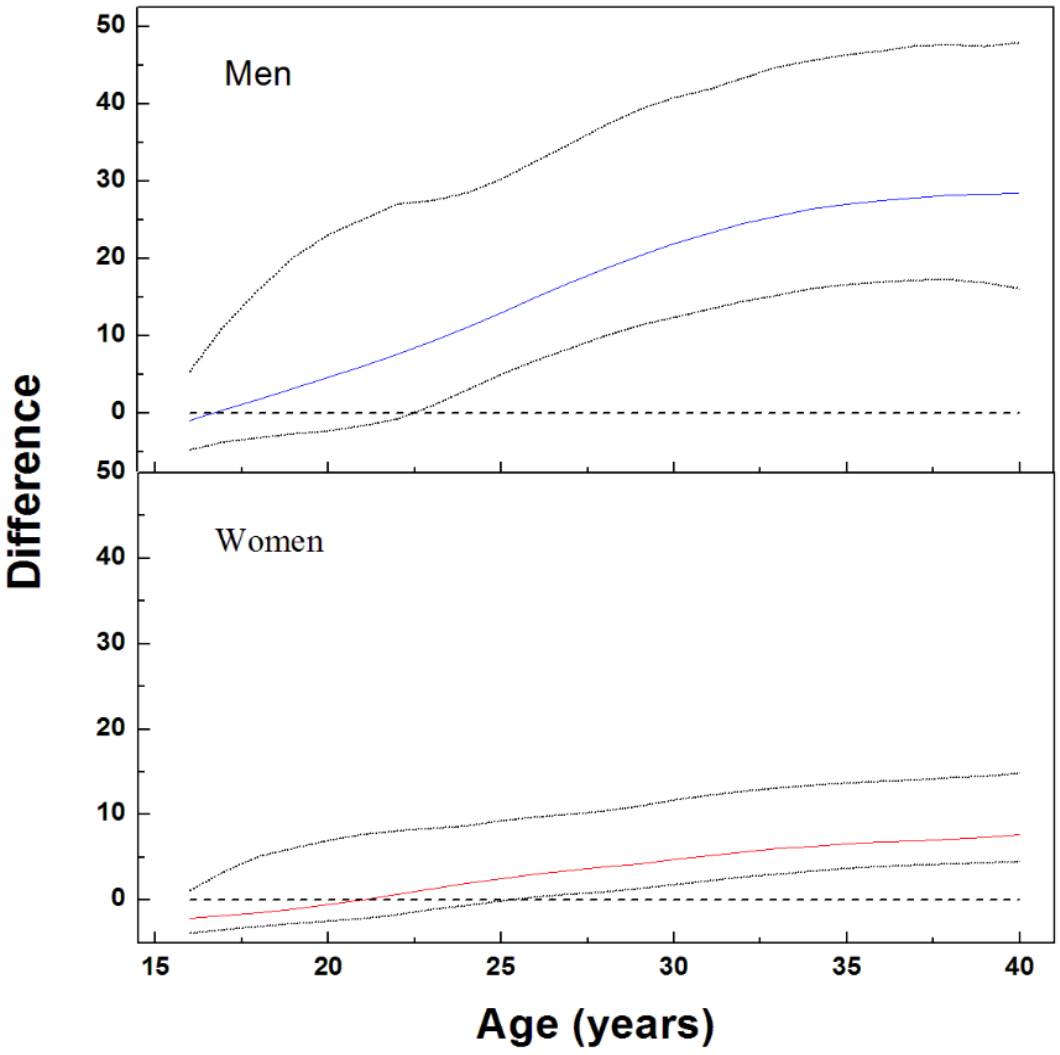


Figure 5.7: The *top panel* shows the confidence band around the bias (blue line) in the synthetic cohort and self-reported data for the males, and the *bottom panel* shows the confidence band around the bias (red line) in the synthetic cohort and self-reported data for the females, using the bootstrap method. Please note that bias results from the difference between the expected values and the self - reported values as observed in the data.

We estimated the confidence interval estimates around it by constructing 1000 bootstrap replications for each gender stratum. The difference in the lifetime number of sexual partners obtained from the self-reported data and the synthetic cohort approach was constructed and plotted in Figure 5.7. The confidence intervals around both figures do not increase that much in width as a function of age.

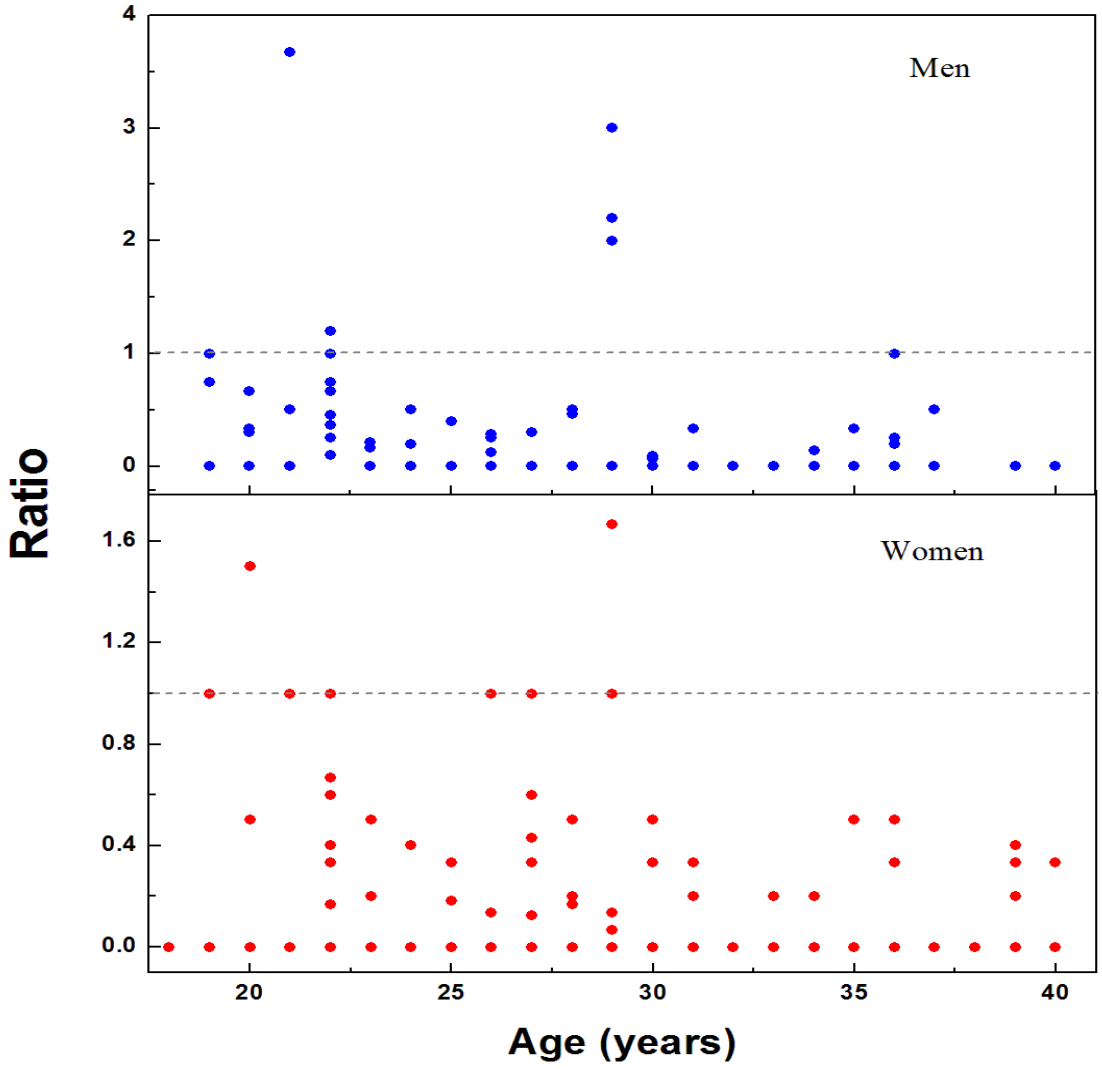


Figure 5.8: In this figure, the *top panel* shows the ratio of the number of new sexual partners formed in the last year and the lifetime number of sexual partners for the men and the *bottom panel* shows for the women. The values above one in both plots show untrue information because the number of new sexual partners in the last year should not be greater than the lifetime number of sexual partners.

The ratio of the number of new sexual partners in the last year (NPLY) and the lifetime number of sexual partners (LNP) are shown in Figure 5.8 as shown in both panels in the figure above. Some values are above one and this indicates that the individuals gave an untrue information.

5.6 Predictive power of risk factors on HIV status

Table 5.1 presents the HIV prevalence of our sample by age and gender. In age group 16 to 20 years, 16% of the women were HIV positive and none of the men tested positive. In all the age groups, women had a higher percentage of being HIV positive than the men.

Table 5.25: HIV prevalence by age and gender (in %)

	Men	**Women**
16-20 years	0	16
21-25 years	4	36
26-30 years	10	50
31-35 years	23	55
36-40 years	36	47

5.6.0.1 Test for multicollinearity

From Chapter 4, section 4.6, our models were:

$$\text{Model 1: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3,$$

$$\text{Model 2: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3 + \beta_4 X_{2i},$$

$$\text{Model 3: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3 + \beta_4 X_{3i},$$

$$\text{Model 4: } f(\log[\pi(y_i)]) = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - t_1)^3 + \beta_3 (X_{1i} - t_2)^3 + \beta_4 X_{2i} + \beta_5 X_{3i},$$

where X_{1i} is the age of subject i, $\pi(y_i)$ is subject i's underlying risk of having HIV, X_{2i} is new number of sexual partners in the last year for subject i and X_{3i} is lifetime number of sexual partners for subject i.

To ensure that the new number of sexual partners in the last year preceding the survey and the lifetime number of sexual partners are not correlated, a multicollinearity test was carried out. This was done by calculating the variance inflation factor (VIF)

for each model, except for Model 1, which contains fewer than two predictors. Multicollinearity exists if the value of the VIF > 10.

Table 5.26: VIF for models in the male population

	Model 2	Model 3	Model 4
Age	1.28	1.02	1.30
NPLY	1.28	1.02	1.29
LNP			1.04

Table 5.27: VIF for models in the female population

	Model 2	Model 3	Model 4
Age	1.08	1.07	1.15
NPLY	1.08	1.07	1.13
LNP			1.12

None of the variance inflation factors (VIF) are > 10. Thus, multicollinearity does not exist and the variables can be included in the analysis.

5.6.0.2 Results from the modified poisson regression model

It is important to note that age was allowed to be non-linear and $\mathbf{k} = 3$ was used (see section 2.7.4). The estimates obtained from Model 2 (for men) and Model 4 (for women) are shown below (Tables 5.28 and 5.29). Estimates from Models 1, 3 and 4 for men and Models 1 to 3 for women are shown in Appendix A.

Table 5.28: Estimates from Model 2 for the male population

	Estimate	Std. Error	z value	Pr($>$ |z|)
(Intercept)	-13.4201	4.6245	-2.9019	0.0037
ns(age, 3)1	6.1747	2.9133	2.1194	0.0341
ns(age, 3)2	22.9402	9.1721	2.5011	0.0124
ns(age, 3)3	6.0689	1.9913	3.0477	0.00230
NPLY	0.1098	0.1209	0.9082	0.3637

When subjected to cross validation, Models 2, 3 and 4 yielded the lowest prediction error for the men but we selected model 2 as our preferred model because of its simplicity. For the women, Model 4 was selected as the preferred model (see Table 5.30).
After adjusting for the effect of age in the male population, the covariates added in

Table 5.29: Estimates from Model 4 for the female population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.2155	0.5631	-3.9345	8.335e-05
ns(age, 3)1	1.0260	0.4061	2.5267	0.0115
ns(age, 3)2	2.1204	1.2126	1.7486	0.0804
ns(age, 3)3	0.6281	0.2939	2.1369	0.0326
NPLY	0.1712	0.0810	2.1124	0.0347
LNP	0.0645	0.0221	2.9181	0.0035

Model 1 yielded a small prediction error (17%) which means that predictions from Model 1 is different from the true values by 17% error rate. Models 2, 3 and 4 respectively yielded smaller but same prediction errors (16%), which shows that the other covariates added in Models 3 and 4 do not have a predictive effect on HIV status, albeit Model 2 is the simplest model of the three and hence it was chosen. After adjusting for the effect of age in the black women population, Models 1, 2 ,3 and 4 successively yielded smaller prediction errors, hence Model 4 (39%) was chosen (see Table 5.30). This means that the predictions from Model 4 is different from the true values by 39% error rate. However, we need to be sure that the differences between the prediction error estimates obtained were statistically significant, thus the bootstrap method was used to obtain the confidence intervals around these differences as shown in Table 5.31. The confidence intervals indicate that the differences are not statistically significant and that a model is no better than the other. Merely looking at the differences in the prediction errors, they were too small to conclude that the total number of new sexual partners in the last year (NPLY) and the lifetime number of sexual partners (LNP) have predictive effect on HIV status. This study concludes that these indicators are not good predictors of HIV status.

Table 5.30: Table reporting prediction errors from models (in %)

Models	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Men	17.2	15.6	15.6	15.6
Women	48.1	46.1	42.7	39.3

5.7 Summary

The results revealed a number of important aspects. Age and gender have an association with the number of new sexual partners in the last year and the lifetime number of

Table 5.31: Differences between prediction error estimates and their 95% confidence intervals (in %)

	Models 1 & 2		**Models 1 & 3**		**Models 1 & 4**	
	Difference	95 % C.I	Difference	95 % C.I	Difference	95 % C.I
Men	1.6	-4.7 – 7.8	1.6	-6.3 – 7.8	1.6	-6.3 – 10.9
Women	2.0	-4.9 – 18.9	5.4	-4.4 – 16.5	8.8	-3.9 – 19.4

sexual partners. Some inconsistencies were found in the reported and the expected lifetime number of sexual partners. Finally, after adjusting for the effect of age and gender, the study found these indicators, number of new sexual partners in the last year and the lifetime number of sexual partners have very low predictive power on HIV status.

The next chapter will give a discussion of the results. We will go on to draw conclusions and then present the limitations of the study as well as suggestions for further research.

Chapter 6

Discussion and Conclusion

6.1 Introduction

The data presented in this study was conducted in three communities in Western Cape. We aimed to explore the effect of age and gender on the number of sexual partners (number of new sexual partners in the last year preceding the survey and the lifetime number of sexual partners), and also to see if the number of sexual partners is a good predictor of HIV status (Adimora *et al.*, 2007; Catania *et al.*, 2005). This study focused on three aspects, namely, the effect of age and gender on the number of sexual partners, inconsistencies in the reported and expected lifetime number of sexual partners, and investigating if the number of new sexual partners in the past year and the lifetime number of sexual partners is predictive of HIV status. The results confirmed past research (Cubbins and Tanfer, 2000; Zimmer-Gembeck and Collins, 2008; Wittrock, 2004) by showing an effect of age and gender on the number of sexual partners. It also shows that the reported and expected lifetime number of sexual partners are inconsistent. Lastly, it reveals that the number of new sexual partners in the past year (12 months preceding the survey) and the lifetime number of sexual partners are not strong enough to predict HIV status. The results of this data analysis will be discussed and compared to previous findings of research. The conclusion of the present study will be presented and the limitations with further recommendations will be made for further research.

6.2 Discussion of findings

The cubic spline method was used because we want the real effect of age on the variables of interest to be clearly seen. The natural cubic spline method was used in order

to control the boundaries of the curve. We do not want the curves to be distorted at the ends.

The results presented in section 5.4.2 illustrates the Poisson regression estimates for both the male and female population. The Poisson regression exhibits overdispersion, which shows that heterogeneity is present in our data. The next step was to use a negative binomial regression approach, which controls for overdispersion in the data. The estimates are presented in section 5.4.4. A comparative exercise was done on the results presented by the Poisson and negative binomial approach as displayed in section 5.4.5. The estimates from both regression approaches were compared using parameter coefficients, standard error values and the 95% confidence interval estimates.

The parameter coefficients obtained are similar to each other. This is not surprising as both have identical underlying distributions, except for the extra overdispersion term present in the negative binomial distribution. The standard error values are higher in the negative binomial regression approach because the overdispersion in the analysis would cause the errors to be underestimated in a Poisson model. Thus, for the two models presented, the Poisson standard error estimates are biased, but lower than the negative binomial regression estimates.

Confidence intervals give us a sense of how precise the regression coefficients are. A 95% confidence interval means that regardless of the number of time the experiment is conducted, the true value is in that interval 95% of the time. The 95% confidence interval estimates for the negative binomial regression are wider than the Poisson regression estimates.

6.2.1 Findings related to the effect of age and gender

The effect of age was found to be non-linear as each age has different behaviour in acquiring new sexual partners in the year preceding the survey and also in their lifetime. This was obtained by the use of the natural cubic spline. Using the cubic spline method has helped to detect the differing behaviours by age. This shows that in doing a regression analysis on data obtained from behavioural studies, age should be made flexible so as to reveal the non-linear effects, if there are.

For the gender effect, the Poisson regression model was initially used to model the relationship between the number of new sexual partners in the past year preceding the survey, lifetime number of sexual partners, age and gender, but an evidence of overdispersion was found. The Poisson model was used initially because it is the benchmark model for count data (see section 2.5.1). This led us to use the negative binomial regression , which accounts for overdispersion.

The negative binomial regression model results showed a great variation between the number of sexual partners reported by the men and women (see chapter 5, Figure 5.5). This supports the findings of Smith (1992) who pointed out that on average, men report a higher number of sexual partners than women. Various studies have shown that when it comes to sexual behaviour data, men are inclined to over-report while women under-report for various reasons.

Firstly, a man may be compelled to exaggerate if he feels that he should have a lot of sexual partners, and if a woman also feels she should have fewer partners, she will be compelled to under-report (Jonason and Fisher, 2009; Haavio-Mannila and Roos, 2007). It could also be the case that men measures their own status in terms of the number of sexual partners they have had. Aside from these factors, it could be that both men and women have different views and definitions of what having a sex partner means. These are the possible reasons why there is a difference between the reported number of sexual partners for men and women Wellings and Mitchell (2012).

Also, the younger birth cohorts have reported a higher number of sexual partners than their older counterparts (see chapter 5, Figures 5.1 and 5.2). This agrees with the findings of Stigum *et al.* (1997) who found that individuals in the younger birth cohort reported higher lifetime number of sexual partners than the older birth cohort

This may be due to generational differences, and people get more exposed to risky sexual behaviours with time Smith (1992). A rapid change in technology, which has increased diversity in social networks or change in sexual beliefs could also be a possible cause for this.

Usage of alcohol and drugs increases the risk of early sexual intercourse, which could also be a contributing factor to having multiple partners Haavio-Mannila and Roos (2007). This is validated by the findings of Rosenbaum et al (Rosenbaum and Kandel, 1990; Santelli *et al.*, 1998).

Religious and cultural beliefs could be other reasons due to the fact that the older birth cohort abided by the religious and cultural laws against having a sexual partner before marriage (MARRI, 1992; Fagan, 2008). These are the reasons for the younger cohort to have more number of sexual partners than their old counterparts.

6.2.2 Inconsistencies between the reported and expected lifetime number of partners

The synthetic cohort approach was used in estimating the expected lifetime number of sexual partners for each age and this is a feasible method to use. A key advantage of this approach is that it allowed us to study a long period of time using the reports given

by the data. The synthetic cohort approach does not give a proper representation of the sample from the population of people born in a certain year and the people who are still alive and can remember the exact number of sexual partners they have had in their lifetime. People who are very old at the time the survey was held may likely not remember their lifetime number of partners. This was solved by using respondents younger than 40 years old.

The reported and expected lifetime number of sexual partners were found to exhibit great discrepancies (see chapter 5, Figure 5.6). This is close to the findings of Brewer et al Brewer *et al.* (2000) who discovered that the number of sexual partnerships reported by female sex-workers in national surveys is inconsistent with the number expected.

One of the reasons for the discrepancies could be that the participants gave false information. The fear of being watched or traced could make participants give reports that are not true and this will introduce bias into our results Fenton *et al.* (2001).

Another reason could be forgetfulness (recall bias) Fenton *et al.* (2001). This is very likely when an individual has a high rate of partner change and it is easy to forget the total number of sexual partners he has had in his lifetime. If this data were used to predict the HIV status of this individual, there is a high probability that misclassification will occur - the individual may be grouped as HIV negative instead of positive and vice versa.

The inconsistencies are more evident in the male population due to the fact that men over-report and women under-report Smith (1992). This part of the analysis was carried out because the number of sexual partners is frequently used as a predictor for HIV infection. This finding shows that these indicators may not be accurate for predicting HIV infection.

6.2.3 Findings related to predictive power

The modified Poisson regression model was used to check for the predictive power of the number of new sexual partners in the past year preceding the survey (NPLY) and the lifetime number of sexual partners (LNP) on HIV status after adjusting for the effect of age and gender. This is one feasible method to use instead of the logistic regression model as we can estimate the risk ratio directly.

One would expect that the new number of sexual partners in the past year preceding the survey and the lifetime number of sexual partners people have had are correlated. In order to be sure that this correlation is not so huge that it affects the parameter estimates, a multicollinearity test was carried out and the results are shown in Tables 5.26 and 5.27. The output shows that the variance inflation factor for the predictors age, new number of sexual partners in the past year preceding the survey and the lifetime number of sex-

ual partners are moderately correlated. These values indicate that the correlations are small and not to be worried about.

From our findings (see Table 5.25), women in the early birth cohort got exposed to HIV infection earlier than their male counterparts. This result supports the findings that women are more susceptible to infection Dellar *et al.* (2015) even at a young age which could be due to gender-based violence, and transactional relationships, amongst many other factors.

Multiple sexual partners in terms of a high number of new sexual partners in the past year preceding the survey, and high lifetime number of sexual partners, are believed to be risk factors for HIV Berry and Hall (2011). A new study by Dubbinks et al Dubbink *et al.* (2016) reported an association between age, high number of lifetime sexual partners and HIV infection. Note that this is somewhat different from what we have done in this study because we investigated the effect of lifetime partners on HIV status after adjusting for the effect of age.

From our findings, the total number of new sexual partners in the last year and the lifetime number of sexual partners had little or no predictive power on HIV status. The model with the number of new sexual partners in the last year (see Table 5.30) had the same predictive power as the model with the lifetime number of sexual partners for the male cohort. Although, Todd et al. reported that the number of new sexual partners in the past year may be a more valid indicator than the lifetime number of sexual partnerships Todd *et al.* (2009), this does not seem to be so in our study (see Tables 5.30 and 5.31).

Todd et al. concluded that the number of sexual partners is inadequate to explain differences in HIV status. They suggested that information about the duration and types of partnerships should be included in the analysis Todd *et al.* (2009).

6.3 Future directions and recommendations

Other types of generalized linear models should be used, for instance, the Zero Inflated Models. This is because many people reported zero number of new sexual partners in the past year preceding the survey. These models will accurately account for this.

Also, Agent-based simulation methods can be used to mimic the sample population and then be used to predict the lifetime number of sexual partners instead of a synthetic cohort approach. This can bring about a different perspective of the indicators we examined.

6.4 Conclusions

The objective of this study was to investigate the indicators, number of new sexual partners in the last year, and the lifetime number of sexual partners, as good predictors of HIV acquisition risk, by using sample from three underprivileged communities in Cape Town, South Africa. The study used qualitative research methods to gain insight into the number of new sexual partners people had in the last year preceding the survey, the lifetime number of sexual partners and HIV status of the respondents. The study investigated black males and females between the ages of 16 and 40.

The present study found a non-linear effect of age on the number of new sexual partners in the last year and the lifetime number of sexual partners, using a natural cubic spline regression approach. This approach allows us to model non-linear relationships between a response variable of interest, and the covariates being examined. On average, people in the younger age cohorts reported more sexual partners than their older counterparts. In the physical world rather than the mathematical world, situations/relationships/behaviours are usually not continuous (i.e disjoint) in nature. In our own case, the behaviour in one age group is different from the behaviour in another age group. This cannot be captured well by the polynomial functions, as splines have smoother curves to illustrate the physical world.

Our study further revealed inconsistencies between the reported and the expected lifetime number of sexual partners. The extent to which the number of new sexual partners and the lifetime number of sexual partners are predictive of HIV status was investigated in this study using the modified Poisson regression model.

Four models were built and the leave-one-out cross validation (LOOCV) method was used to select the best feasible model in terms of its predictive power on HIV status.

It could not be concluded from our findings that these indicators have predictive power on HIV status. A possible explanation for this could be that participants gave false information about the number of new sexual partners and the lifetime number of sexual partners, this could introduce bias into the results. Also, due to long reporting periods, participants with high partner turnover rate may have forgotten their number of sexual partners in the past. This leads to under-reporting or exaggeration, which may introduce bias into the study. It could also be the case that foreign nationals were included in the data as it was not related to citizenship or residency. These foreign nationals could have come from low risk areas and are not highly exposed to the risk of acquiring HIV, yet they have had a high partner turnover rate. Moreover, it could be the case that the total number of new sexual partners in the last year and the lifetime number of sexual

partners had little or no predictive power on HIV status because the relationship is not strong. These indicators may not have much information about HIV status. Further research should take this into account as they are liable to give unreliable results. These limitations can be used as a pathway for further research.

Appendix A

A.1 Estimates from models

Below are the estimates obtained from the modified Poisson regression models in section 4.5 (page 41).

Table A.1: Estimates from Model 1 for the male population

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-14.4062	5.8334	-2.4696	0.01353
ns(age, 3)1	6.9507	3.6961	1.8805	0.06003
ns(age, 3)2	25.1365	11.3094	2.2226	0.02624
ns(age, 3)3	6.3238	2.5183	2.5112	0.01203

Table A.2: Estimates from Model 1 for the female population

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-2.0354	0.5626	-3.6181	0.0003
ns(age, 3)1	0.9819	0.3937	2.4943	0.0126
ns(age, 3)2	2.4979	1.2042	2.0743	0.0380
ns(age, 3)3	0.5402	0.3010	1.7943	0.0728

Table A.3: Estimates from Model 2 for the female population

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-2.1586	0.5891	-3.6640	0.0003
ns(age, 3)1	1.1510	0.4108	2.8019	0.0051
ns(age, 3)2	2.5316	1.2382	2.0446	0.0409
ns(age, 3)3	0.6069	0.3071	1.9763	0.0481
NPLY	0.2134	0.0787	2.7123	0.0067

Table A.4: Estimates from Model 3 for the male population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-14.4069	5.9064	-2.4392	0.0147
ns(age, 3)1	7.0051	3.7321	1.8770	0.0605
ns(age, 3)2	25.3449	11.3668	2.2297	0.0258
ns(age, 3)3	6.3417	2.5479	2.4890	0.0128
LNP	-0.0190	0.0648	-0.2941	0.7687

Table A.5: Estimates from Model 3 for the female population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.1497	0.5460	-3.9370	8.249e-05
ns(age, 3)1	0.8804	0.3881	2.2686	0.0233
ns(age, 3)2	2.0779	1.1860	1.7520	0.0798
ns(age, 3)3	0.5862	0.2881	2.0347	0.0419
LNP	0.0743	0.0213	3.4858	0.0005

Table A.6: Estimates from Model 4 for the male population

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-13.3346	4.6675	-2.8569	0.0043
ns(age, 3)1	6.1809	2.9269	2.1117	0.0347
ns(age, 3)2	23.0464	9.1769	2.5114	0.0120
ns(age, 3)3	6.0535	2.0225	2.9930	0.0028
NPLY	0.1133	0.1111	1.0199	0.3078
LNP	-0.0252	0.0686	-0.3667	0.7139

Appendix B

B.1 Investigating inconsistencies between the reported and predicted lifetime number of sexual partners in Cape Town, South Africa

In 2011, South Africa (SA) was one of the countries with the world's highest prevalence of HIV/AIDS UNAIDS (2013). Common indicators such as the number of new sexual partners in a given year and the lifetime number of sexual partners are used in several analyses to predict the risk of contracting HIV. However, are these indicators consistent?

A cross-sectional sexual behaviour study was conducted using a touch screen questionnaire that utilized an audio computer-assisted self-interviews application (ACASI) to obtain sexual history data. The study was conducted in three disadvantaged communities - a predominantly black community, and other two racially diverse communities which consist of black Africans and the coloured population. These two races are investigated because they have the highest prevalence of HIV in SA Shisana *et al.* (2012).

We performed a negative binomial regression analysis of the number of new sexual partners reported in the year before the survey, as a function of age. For the purpose of this study, the 16-40 year old population was divided into sub-population by race and gender: black men, black women, coloured men and coloured women. The analysis was conducted separately for each sub-population. A synthetic cohort approach was used to estimate the expected lifetime number of sexual partners based on the reported number of new sexual partners in the last year, and this was compared to the reported lifetime number of sexual partners (Figure 1). A synthetic cohort is a hypothetical cohort of people who would be subject at each age to the age-specific rates (rates of acquiring new sexual partners) of one specific period. In our case, the specific period is the 12-month period in 2011-2012 prior to the cross-sectional survey (the exact calendar time period is not the same for all respondents because they were not all interviewed on the

same day, but variation is minimal).

The median age of the 352 participants was 29 years. About 23% were men and 73% were women. There were 18% black men and 59% black women, 5% were coloured men and 18% were coloured women. The total number of partners last year varied from zero to 11 and the lifetime number of sexual partners varied from one to 15. In Figure 1, the blue line is the predicted number of lifetime partners, derived from the negative binomial regression model, fitted to self-reported lifetime partner data (red line).

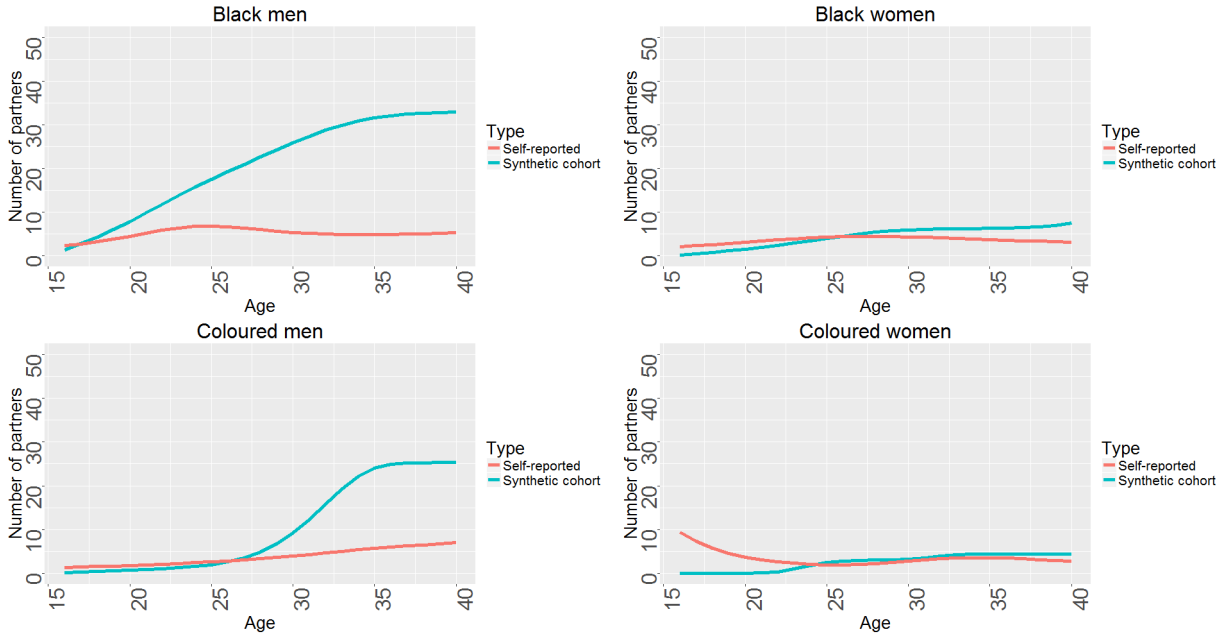


Figure B.1: PREDICTED AND REPORTED LIFETIME NUMBER OF SEXUAL PARTNERS FOR EACH SUB POPULATION

In the sub-population for black men and coloured men, there is a huge discrepancy between the expected lifetime number of partners and the reported lifetime number of partners compared to the women. This may be due to the fact that men over-report and women under-report their number of sexual partners Brown and Sinclair (1999), which may be due to social desirability bias. Sexual behavioural studies are necessary to understand sexual networks and transmission dynamics of HIV. The present study draws attention to the fact that these indicators, number of new sexual partners in the last year and the lifetime number of sexual partners, are inconsistent and therefore inaccurate indicators of sexual risk behaviour especially for men.

Appendix C

C.1 Data analysis

```
data <- data
data.filtered <- dplyr::filter(data,
                               race == 2,
                               gender != "2",
                               age > 15,
                               age < 41)
# We create two subsets, one for each race-gender stratum
BM <- dplyr::filter(data.filtered,
                            race == 2,
                            gender == "0")
BW <- dplyr::filter(data.filtered,
                            race == 2,
                            gender == "1")
##############################################
# Function to bootstrap the data
##############################################
diffest<- function(data,indices){
  d <- data[indices,] # taking a random sample of the full dataset
  final<-glm.nb(totalpartnerslstyr~ns(age,4),
                data=d,
                control=glm.control(maxit=100000, trace = 3),
                init.theta=0.5) # fitting the PLY model to this sample
  final2<-glm.nb(totalpartners~ns(age,4),
                 data=d,
                 control=glm.control(maxit=100000, trace = 3),
```

```
                init.theta = 5) # fitting the LTP model to this sample

  made <- data.frame(age=16:40) # creating artificial dataset for model predictions

  try <- predict(final, newdata=made, type="response") # model predictions for PLY
  s <- cumsum(try) # cumulating PLY to obtain a model-based estimate of LTP

  tryL <- predict(final2, newdata=made, type="response") # model predictions for LTP
 # Difference between the cumulated estimates and the model-prediction estimates
  diffvect <- s - tryL
  return(diffvect)
}
# Model for black men in the last year

finalBm<-glm.nb(totalpartnerslstyr~ns(age,4),
              data=BM,
              control=glm.control(maxit=100000, trace = 3),
              init.theta=0.5)
summary(finalBm)

# Model for black women in the last year

finalBw<-glm.nb(totalpartnerslstyr~ns(age,4),
                data=BW,
                control=glm.control(maxit=100000, trace = 3),
                init.theta=0.5) # fitting the PLY model to this sample
summary(finalBw)

# Model for the lifetime number of partners for black men

final2Bm<-glm.nb(totalpartners~ns(age,4),
                 data=BM,
                 control=glm.control(maxit=100000, trace = 3),
                 init.theta=5)
summary(final2Bm)
```

```
# Model for the lifetime number of partners for black women

final2Bw<-glm.nb(totalpartners~ns(age,4),
                 data=BW,
                 control=glm.control(maxit=100000, trace = 3),
                 init.theta=5) # fitting the LTP model to this sample
summary(final2Bw)

#We create data frames for black men and women
#to predict new number of partners in the last year

madeforBM<-data.frame(age = 16:40, gender = "man", race = "African")
madeforBW<-data.frame(age = 16:40, gender = "woman", race = "African")

# The response variables (new number of partners in the last year)
madeforBM$totalpartnerslstyr <- NA
madeforBW$totalpartnerslstyr <- NA

# We predict the values for the new data using the models built
predBM<-predict(finalBm,newdata=madeforBM,type="response")
predBW<-predict(finalBw,newdata=madeforBW,type="response")

madeforBM$pred<-predBM;madeforBW$pred<-predBW
madeforBM$s<-cumsum(madeforBM$pred)
madeforBW$u<-cumsum(madeforBW$pred)

s<-cumsum(predBM)
u<-cumsum(predBW)

# Do the same for lifetime number of partners
madeforBML <- data.frame(age = 16:40, gender = "man", race = "African")
madeforBWL <- data.frame(age = 16:40, gender = "woman", race = "African")

# The response variables
madeforBML$totalpartners <- NA
madeforBWL$totalpartners <- NA
```

```
# We predict the values for the new data using the models built
# model predictions for LNP
predBML <- predict(final2Bm,newdata=madeforBML,type="response")
predBWL <- predict(final2Bw,newdata=madeforBWL,type="response")
madeforBML$predL <- predBML;madeforBWL$predL<-predBWL

madeforBML$s <- s
madeforBWL$u <- u

madeforBML$YL <- (madeforBML$s - madeforBML$predL)
madeforBWL$YL <- (madeforBWL$u - madeforBWL$predL)
# Difference between the cumulated estimates and the model-prediction estimates
diffvectBM <- s-predBML
diffvectBW <- u-predBWL
# Contains the values resulting from the bootstrap
diffvectALL <- c(diffvectBM, diffvectBW)

#########################################################
# We construct confidence band around our values.
# We do this 2 times: once for each stratum

# 1. For Men
results<-boot(data=BM, statistic = diffest, R=1000)
lowerbandBM <-rep(NA,length(results$t0))
upperbandBM <- lowerbandBM
for (i in 1:length(results$t0) ) {
  BCI <- boot.ci(results, type="perc", index=i)
  lowerbandBM[i] <- BCI$percent[4]
  upperbandBM[i] <- BCI$percent[5]
}

# 2. For Women
resultsBW<-boot(data=BW, statistic = diffest, R=1000)
lowerbandBW <-rep(NA,length(resultsBW$t0))
upperbandBW <- lowerbandBW
```

```
for (i in 1:length(resultsBW$t0) ) {
  BCI.BW <- boot.ci(resultsBW, type="perc", index=i)
  lowerbandBW[i] <- BCI.BW$percent[4]
  upperbandBW[i] <- BCI.BW$percent[5]
}
### Final step: plotting the results
plot(seq(16,40,length=25),diffvectBM, type="l",
col="navyblue",main="Black men",xlab="Age",ylab="Difference",ylim=c(-20,75))
lines(seq(16,40,length=25),lowerbandBM[1:25], lty=2)
lines(seq(16,40,length=25),upperbandBM[1:25], lty=2)
lines(c(16,40), c(0,0), lty=3)

plot(seq(16,40,length=25),diffvectBW, type="l",
col="red3",main="Black women",xlab="Age",ylab="Difference",ylim=c(-10,25))
lines(seq(16,40,length=25),lowerbandBW, lty=2)
lines(seq(16,40,length=25),upperbandBW, lty=2)
lines(c(16,40), c(0,0), lty=3)

##########################################
#Predictive power for HIV
##########################################

# data.filtered.forHIV is the dataset for the HIV status analysis

data.filtered.forHIV <- dplyr::filter(data.filtered,
                                      hivtestresult < 2)
BMhiv<- dplyr::filter(data.filtered.forHIV,
                                    gender == "0")
BWhiv<- dplyr::filter(data.filtered.forHIV,
                                gender == "1")

md.wideBMhiv$hivtestresult <-as.numeric(md.wideBMhiv$hivtestresult)
md.wideBWhiv$hivtestresult <-as.numeric(md.wideBWhiv$hivtestresult)
```

```
#####################################################################
## Do this for each stratum
## These are the models (Modified Poisson regression models)
#####################################################################

Model1_Bm<-glm(formula = hivtestresult ~ ns(age,3),   #Our model
           data    = BMhiv[complete.cases(BMhiv), ],
           family  = poisson(link = "log"))
Model1_Bw<-glm(formula = hivtestresult ~ ns(age,4),
           data    = BWhiv[complete.cases(BWhiv), ],
           family  = poisson(link = "log"))
Model2_Bm<-glm(formula = hivtestresult ~ ns(age,3) + totalpartnerslstyr,
            data    = BMhiv[complete.cases(BMhiv), ],
            family  = poisson(link = "log"))
Model2_Bw<-glm(formula = hivtestresult ~ ns(age,3) + totalpartnerslstyr,
            data    = BWhiv[complete.cases(BWhiv), ],
            family  = poisson(link = "log"))
Model3_Bm<-glm(formula = hivtestresult ~ ns(age,3) + totalpartners,
            data    = BMhiv[complete.cases(BMhiv), ],
            family  = poisson(link = "log"))
Model3_Bw<-glm(formula = hivtestresult ~ ns(age,3) + totalpartners,
            data    = BWhiv[complete.cases(BWhiv), ],
            family  = poisson(link = "log"))
Model4_Bm<-glm(formula = hivtestresult ~ ns(age,3) +
totalpartnerslstyr + totalpartners,
            data    = BMhiv[complete.cases(BMhiv), ],
            family  = poisson(link = "log"))
Model4_<-glm(formula = hivtestresult ~ ns(age,3) +
totalpartnerslstyr + totalpartners,
            data    = md.wideBWhiv[complete.cases(md.wideBWhiv), ],
            family  = poisson(link = "log"))

#####################################################
#Cross validation
#####################################################
```

```
looc<-function(model){
#counts the no of row in the data associated with the model
  n<- nrow(model$data)
  #creates an empty vessel for predicted values
  yhat<- rep(NA,n)
  for(i in 1:n){ # our k=n (Leave-one out)
    yhat[i] <- predict(update(model, data =model$data[-i, ]),
    model$data[i,],type="response")
  } #predicts the variable of interest using the trained data set.
  #It loops over each training data set because this is the leave-one-out
  pred.values<- yhat #our predicted values for the variable of interest
  data<-model$data  #We store our data into a vector called data
  delta.estimate<-sum(round(pred.values) !=
  data$hivtestresult[complete.cases(data)])/nrow(data)
  # it compares the predicted values with the original data
  #and calculates the error percent
  #returns the pred error
  return(list(as.data.frame(pred.values),delta.estimate=delta.estimate))
  #estimate (delta.estimate) and the data frame
  #that contains the predicted values (pred.values)
}

loc1M<-looc(Model1_Bm) #run the function over the model

#To check if the difference in the prediction error of the
#estimates are not just due to chance, we made a bootstrap
#replication of the differences in the prediction error estimates
#between the models.

boot_PE.diff <- function(dat, index){
  resamp.dat <- dat[index, ]  #resampled data
  Mode1_Bm<-glm(formula = hivtestresult ~ ns(age,3),
              data    = resamp.dat,
              family  = poisson(link = "log"),
              control=glm.control(maxit=100000, trace = 3),
```

```
            init.theta=0.5)
  Model2_Bm<-glm(formula = hivtestresult ~ ns(age,3) + totalpartnerslstyr,
              data    = resamp.dat,
              family  = poisson(link = "log"),
              control=glm.control(maxit=100000, trace = 3),
              init.theta=0.5)
  Model3_Bm<-glm(formula = hivtestresult ~ ns(age,3) + totalpartners,
              data    = resamp.dat,
              family  = poisson(link = "log"),
              control=glm.control(maxit=100000, trace = 3),
              init.theta=0.5)
  Model4_Bm<-glm(formula = hivtestresult ~ ns(age,3) +
  totalpartnerslstyr + totalpartners,
              data    = resamp.dat,
              family  = poisson(link = "log"),
              control=glm.control(maxit=100000, trace = 3),
              init.theta=0.5)
  # predict the values on the resampled data and calculate the error
  vec1 <- predict(MOdel1_Bm, newdata = resamp.dat,type="response") -
  predict(Model2_Bm, newdata = resamp.dat,type="response")
  vec2 <- predict(Model1_Bm, newdata = resamp.dat,type="response") -
  predict(Model3_Bm, newdata = resamp.dat,type="response")
  vec3 <- predict(Model1_Bm, newdata = resamp.dat,type="response") -
  predict(Model4_Bm, newdata = resamp.dat,type="response")
  results <- c(vec1,vec2,vec3)
  return(results)
}

resultsMen <- boot(data=BMhiv[complete.cases(BMhiv), ], statistic=boot_PE.diff,
              R=1000)
              lowerbandBM <-rep(NA,length(resultsMen$t0))
upperbandBM <- lowerbandBM
for (i in 1:length(resultsMen$t0) ) {
  BCIz <- boot.ci(resultsMen, type="perc", index=i)
  lowerbandBM[i] <- BCIz$percent[4]
  upperbandBM[i] <- BCIz$percent[5]
```

```
}

plot(seq(1,64,length=64),vec1, type="l", col="green",
main="Black men",xlab="Individual",ylab="Error",ylim=c(-2,2))
lines(seq(1,64,length=64),lowerbandBM[1:64], lty=2)
lines(seq(1,64,length=64),upperbandBM[1:64], lty=2)
lines(c(1,64), c(0,0), lty=3)

plot(seq(1,64,length=64),vec2, type="l", col="red",
main="Black men",xlab="Individual",ylab="Error",ylim=c(-2,2))
lines(seq(1,64,length=64),lowerbandBM[65:128], lty=2)
lines(seq(1,64,length=64),upperbandBM[65:128], lty=2)
lines(c(1,64), c(0,0), lty=3)

plot(seq(1,64,length=64),vec3, type="l", col="blue",
main="Black men",xlab="Individual",ylab="Error",ylim=c(-2,2))
lines(seq(1,64,length=64),lowerbandBM[129:192], lty=2)
lines(seq(1,64,length=64),upperbandBM[129:192], lty=2)
lines(c(1,64), c(0,0), lty=3)
################
#Women

Model1_Bw<-glm(formula = hivtestresult ~ ns(age,3),
           data    = BWhiv[complete.cases(BWhiv), ],
           family  = poisson(link = "log"),
           control=glm.control(maxit=100000, trace = 3),
           init.theta=0.5)
Model2_Bw<-glm(formula = hivtestresult ~ ns(age,3) + totalpartnerslstyr,
            data    = BWhiv[complete.cases(BWhiv), ],
            family  = poisson(link = "log"),
            control=glm.control(maxit=100000, trace = 3),
            init.theta=0.5)
Model3_Bw<-glm(formula = hivtestresult ~ ns(age,3) + totalpartners,
            data    = BWhiv[complete.cases(BWhiv), ],
            family  = poisson(link = "log"),
            control=glm.control(maxit=100000, trace = 3),
```

```
            init.theta=0.5)
Model4_Bw<-glm(formula = hivtestresult ~ ns(age,3) +
totalpartnerslstyr + totalpartners,
            data    = BWhiv[complete.cases(BWhiv), ],
            family  = poisson(link = "log"),
            control=glm.control(maxit=100000, trace = 3),
            init.theta=0.5)
vec1 <- predict(Model1_Bw, newdata = BWhiv[complete.cases(BWhiv), ],
type="response") - predict(Model2_Bw, newdata = BWhiv[complete.cases(BWhiv), ],
  type="response")
vec2 <- predict(Model1_Bw, newdata = BWhiv[complete.cases(BWhiv), ],
type="response") - predict(Model3_Bw, newdata = BWhiv[complete.cases(BWhiv), ],
type="response")
vec3 <- predict(Model1_Bw, newdata = BWhiv[complete.cases(BWhiv), ],
type="response") - predict(Model4_Bw, newdata = BWhiv[complete.cases(BWhiv), ],
type="response")

resultsWomen <- boot(data=BWhiv[complete.cases(BWhiv), ], statistic=boot_PE.diff,
                R=1000)

lowerbandBW <-rep(NA,length(resultsWomen$t0))
upperbandBW <- lowerbandBW
for (i in 1:length(resultsWomen$t0) ) {
  BCIz <- boot.ci(resultsWomen, type="perc", index=i)
  lowerbandBW[i] <- BCIz$percent[4]
  upperbandBW[i] <- BCIz$percent[5]
}
```

List of references

Adimora, A.A., Schoenbach, V.J. and Doherty, I.A. (2007). Concurrent sexual partnerships among men in the United States. *American Journal of Public Health*, vol. 97, no. 12, pp. 2230–2237.

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons Ltd, United States of America.

Ahlberg, J.H., Nilson, E.N. and Walsh, J.L. (1967). *The theory of splines and their applications*. Academic Press Inc., New York and London.

Alexander, L.K., Lopes, B., Ricchetti-Masterson, K. and Yeatts, K.B. (2015). *Confounding bias , part II and effect measure modification*. 2nd edn. The University of North Carolina at Chapel Hill, Department of Epidemiology.

Anderson, R.M. and May, R.M. (1988). Epidemiological parameters of hiv. *Nature*, vol. 333.

Arora, P., Nagelkerke, N.J. and Jha, P. (2012). A systematic review and meta-analysis of risk factors for sexual transmission of HIV in India. *PLoS ONE*, vol. 7, no. 8.

Atkinson, K.E. (1989). *An introduction to numerical analysis*. John Wiley & Sons, Inc., Canada.

Attanasio, O.P. (1993). An analysis of life-cycle accumulation assets. *Ricerche Economiche*, vol. 47, no. 4, pp. 323 – 354.

Ayles, H.M., Sismanidis, C., Beyers, N., Hayes, R.J. and Godfrey-faussett, P. (2008). ZAMSTAR , The Zambia South Africa TB and HIV Reduction study: design of a 2 × 2 factorial community randomized trial. *Trials*, vol. 9, no. 1, p. 63.

Barry, M., Dewar, D., Whittal, J. and Muzondo, I. (2007). Land conflicts in informal settlements: Wallacedene in Cape Town, South Africa. In: *Urban Forum*, vol. 18, pp. 171–189. Springer.

Bell, K.M. and Naugle, A.E. (2005). Understanding stay/leave decisions in violent relationships: A behavior analytic approach. *Behavior and Social Issues*, vol. 14, no. 1, p. 21.

Berko, J. (2014). Deaths attributed to heat , cold , and other weather events in the United States , 2006 – 2010. Tech. Rep. 76, National Center for Health Statistics.

Berry, L. and Hall, K. (2011). Multiple sexual partnerships. Tech. Rep., Children's Institute, University of Cape Town.

Borgdorff, M., Barongo, L., Newell, J., Senkoro, K., Deville, W., Velema, J. and Gabone, R. (1994). Sexual partner change and condom use among urban factory workers in northwest Tanzania. *Genitourin Med*, vol. 70, no. 6, pp. 378–383.

Brewer, D.D., Potterat, J.J., Garrett, S.B., Muth, S.Q., Roberts, J.M., Kasprzyk, D., Montano, D.E. and Darrow, W.W. (2000). Prostitution and the sex discrepancy in reported number of sexual partners. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 22, pp. 12385–12388.

Brown, N.R. and Sinclair, R.C. (1999). Estimating number of lifetime sexual partners: Men and women do it differently. *Journal of Sex Research*, vol. 36, pp. 292–297.

Cameron, A.C. and Trivedi, P.K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Econometrics, Applied*, vol. 1, no. 1, pp. 29–53.

Cameron, A.C. and Trivedi, P.K. (2013). *Regression analysis of count data*. Cambridge University Press, Cambridge.

Catania, J.A., Osmond, D., Neilands, T.B., Canchola, J., Gregorich, S. and Shiboski, S. (2005). Methodological challenges in research on sexual risk behavior: Commentary on Schroder et al. (2003a, 2003b). *Annals of Behavioral Medicine*, vol. 29, no. 2, pp. 86–95.

Center for AIDS Prevention Studies 2003 (). Fact Sheet 50E – How do sexual networks affect HIV / STD prevention?

Centers for Disease Control and Prevention (2016). Anal Sex and HIV Risk. Available at: http://www.cdc.gov/hiv/risk/analsex.html

Choi, Y., Ahn, H. and Chen, J.J. (2005). Regression trees for analysis of count data with extra poisson variation. *Computational statistics & data analysis*, vol. 49, no. 3, pp. 893–915.

Choudhry, V., Ambresin, A., Nyakato, V.N. and Agardh, A. (2015). Transactional sex and HIV risks - evidence from a cross-sectional national survey among young people in Uganda. *Global Health Action*, vol. 8, p. 27249.

Clumeck, N., Van de Perre, P., Carael, M., Rouvroy, D. and Nzaramba, D. (2010). Heterosexual promiscuity among African patients with AIDS. *The New England Journal of Medicine*, p. 2016.

Cohen, D. (2000). Poverty and hiv/aids in sub-saharan africa.

Cooper, J.O., Heron, T.E. and Heward, W.L. (2012). *Applied behaviour analysis*. Pearson/Merrill-Prentice Hall Upper Saddle River, NJ.

Crawley, M.J. (2012). *The R book*. John Wiley & Sons Ltd.

Cubbins, L.A. and Tanfer, K. (2000). The influence of gender on sex: a study of men's and women's self-reported high-risk sex behavior. *Archives of Sexual Behavior*, vol. 29, no. 3, pp. 229–257.

Delft (2015). Delft, Cape Town — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Delft. Accessed: November 28, 2016.

Dellar, R.C., Dlamini, S. and Karim, Q.A. (2015). Adolescent girls and young women: key populations for HIV epidemic control. *Journal of the International AIDS Society*, vol. 18, p. 19408.

Delva, W. (2012). Description of the Cape Town Sexual Network Survey Data. Tech. Rep..

Delva, W., Beauclair, R., Welte, A., Vansteelandt, S., Hens, N., Aerts, M., Du Toit, E., Beyers, N. and Temmerman, M. (2011). Age-disparity, sexual connectedness and HIV infection in disadvantaged communities around Cape Town, South Africa: a study protocol. *BMC public health*, vol. 11, no. 1, p. 616.

Delva, W., Meng, F., Beauclair, R., Deprez, N., Temmerman, M., Welte, A. and Hens, N. (2013). Coital frequency and condom use in monogamous and concurrent sexual relationships in Cape Town, South Africa. http://www.sacema.org/node/coital-frequency-and-condom-use-in-monogamous. Accessed: December 01, 2016.

Dubbink, J.H., Van der Eem, L., McIntyre, J.A., Mbambazela, N., Jobson, G.A., Ouburg, S., Morre, S.A., Struthers, H.E. and Peters, R.P. (2016). Sexual behaviour of women in rural South Africa: a descriptive study. *BMC Public Health*, vol. 16, no. 1, p. 557.

Efron, B. and Tibshirani, R.J. (1994). *An introduction to the bootstrap*. CRC press.

Efron, B. and Tibshirani, R.J. (1995). *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule*. Division of Biostatistics, Stanford University.

Epstein, H. (2010). The mathematics of concurrent partnerships and HIV: a commentary on Lurie and Rosenthal. *AIDS and Behavior*, vol. 14, no. 1, pp. 29–30.

Fagan, P.F. (2008). Religious attendance and number of sexual intercourse partners — adolescent girls. Tech. Rep. 4, Family Research Council.

Fenton, K.A., Johnson, A.M., McManus, S. and Erens, B. (2001). Measuring sexual behaviour: methodological challenges in survey research. *Sexually Transmitted Infections*, vol. 77, no. 2, pp. 84–92.

Finer, L.B., Darroch, J.E. and Singh, S. (1999). Sexual partnership patterns as a behavioral risk factor for sexually transmitted diseases. *Family Planning Perspectives*, vol. 31, no. 5, p. 228.

Fisher, H. (2000). Lust, attraction, attachment: Biology and evolution of the three primary emotion systems for mating, reproduction, and parenting. *Journal of Sex Education and Therapy*, vol. 25, no. 1, pp. 96–104.

Fletcher, G.J., Simpson, J.A., Campbell, L. and Overall, N.C. (2012). *The science of intimate relationships*. John Wiley & Sons.

Friedland, G.H. and Klein, R.S. (1987). Transmission of the human immunodeficiency virus. *New England Journal of Medicine*, vol. 317, no. 18, pp. 1125–1135.

Gardner, W., Mulvey, E.P. and Shaw, E.C. (1995). Regression analyses of counts and rates : Poisson , overdispersed Poisson , and negative binomial models. *Psychological Bulletin*, vol. 118, no. 3, pp. 392–404.

GIS, S.. (2013). City of Cape Town – 2011 census suburb Wallacedene. Tech. Rep., Strategic Development Information and GIS Department, Cape Town.

Haavio-Mannila, E. and Roos, J.P. (2007). Why are men reporting more sexual partners than women? *University of Helsinki, Finland*, vol. 8, pp. 123–137.

Haight, F.A. (1967). *Handbook of the Poisson distribution*, vol. 18. John Wiley & Sons, Inc., New York.

Hall, B.H., Griliches, Z. and Hausman, J.A. (1986). Patents and r and d: Is there a lag? *International economic review*, pp. 265–283.

Halli, S.S. and Rao, V. (1992). *Advanced techniques of population analysis*. Springer Science & Business Media,LLC 2009, New York.

Hallman, K. (2009). Gendered socioeconomic conditions and HIV risk behaviours among young people in South Africa. *African Journal of AIDS Research*, vol. 4, no. 1, pp. 37–50.

Hamilton, W.D. and Howard, J. (2012). *Infection, polymorphism and evolution*. Springer Science & Business Media.

Harrell, F.E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer International Publishing. ISBN 978-3-319-19425-7.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin.

Hausman, J.A., Hall, B.H. and Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. Tech. Rep. 17, Cambridge.

Hilbe, J.M. (2012). *Negative binomial regression*, vol. XXXIII. 2nd edn. Cambridge University Press, New York.

Huang, J. (2012). Natural cubic interpolation. http://www.cas.mcmaster.ca/{~}qiao/courses/cs4xo3/tutorials/CubicSpline.pdf. Accessed: April 20, 2017.

Jaggia, S. and Thosar, S. (1993). Multiple bids as a consequence of target management resistance: a count data approach. *Review of Quantitative Finance and Accounting*, vol. 3, no. 4, pp. 447–457.

Jewkes, R., Dunkle, K., Nduna, M. and Shai, N.J. (2012). Transactional sex and HIV incidence in a cohort of young women in the stepping stones trial. *Journal of AIDS and Clinical Research*, vol. 3, no. 5, p. 7.

Johnson, A.M., Mercer, C.H., Erens, B., Copas, A.J., McManus, S., Wellings, K., Fenton, K.A., Korovessis, C., Macdowall, W., Nanchahal, K. *et al.* (2001). Sexual behaviour in britain: partnerships, practices, and hiv risk behaviours. *The Lancet*, vol. 358, no. 9296, pp. 1835–1842.

Johnson, L., Dorrington, R., Bradshaw, D., Pillay-Van Wyk, V. and Rehle, T. (2009). Sexual behaviour patterns in South Africa and their association with the spread of HIV: insights from a mathematical model. *Demographic Research*, vol. 21, no. 11, pp. 289–340.

Jonason, P.K. and Fisher, T.D. (2009). The power of prestige: Why young men report having more sex partners than young women. *Sex Roles*, vol. 60, no. 3-4, pp. 151–159.

Kabiru, C.W., Luke, N., Izugbara, C.O. and Zulu, E.M. (2010). The correlates of HIV testing and impacts on sexual behavior: evidence from a life history study of young people in Kisumu, Kenya. *BMC Public Health*, vol. 10, no. 1, p. 412.

Kagaayi, J., Gray, R.H., Whalen, C., Fu, P., Neuhauser, D., McGrath, J.W., Sewankambo, N.K., Serwadda, D., Kigozi, G. and Nalugoda, F. (2014). Indices to measure risk of HIV acquisition in Rakai, Uganda. *PLoS ONE*, vol. 9, no. 4, p. e92015.

Kak, L., Chitsike, I., Luo, C. and Rollins, N. (2010). Prevention of mother-to-child transmission of HIV / AIDS programmes. *Opportunities for Africa's Newborns*, pp. 113–126.

Kirby, D., Dayton, R., Lengle, K. and Prickett, A. (2012). *Promoting partner reduction: Helping young people understand and avoid HIV risks from multiple partnerships*. Durham North Carolina FHI 360 2012.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, vol. 34, no. 1, pp. 1–14.

Lamphier, P.A. and Welch, R. (2017). *Women in American history: A social, political, and cultural encyclopedia and document Collection [4 volumes]*. ABC-CLIO.

Laumann, E.O. (1994). *The social organization of sexuality: Sexual practices in the United States*. University of Chicago Press.

Lee, J.H., Han, G., Fulp, W. and Giuliano, A. (2012). Analysis of overdispersed count data: application to the human papillomavirus infection in men (him) study. *Epidemiology and Infection*, vol. 140, no. 06, pp. 1087–1094.

Lehmiller J. (2015). Average number of sex partners - Business Insider. Accessed: November 3, 2016.
Available at: http://www.businessinsider.com/average-number-of-sex-partners-2015-4

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M. and Van Hoewyk, J. (2010). The 2006-2010 national survey of family growth: sample design and analysis of a continuous survey. *Vital and health statistics. Series 2, Data evaluation and methods research*, , no. 150, pp. 1–36.

Levin, K.A. (2006). Study design III: Cross-sectional studies. *Evidence-Based Dentistry*, vol. 7, no. 1, pp. 24–25.
Available at: http://www.nature.com/doifinder/10.1038/sj.ebd.6400375

Long, J.S. and Freese, J. (2001). *Regression models for categorical dependent variables using STATA*. Stata Press.

MacPherson, E.E., Sadalaki, J., Njoloma, M., Nyongopa, V., Nkhwazi, L., Mwapasa, V., Lalloo, D.G., Desmond, N., Seeley, J. and Theobald, S. (2012). Transactional sex and HIV: understanding the gendered structural drivers of HIV in fishing communities in Southern Malawi. *Journal of the International AIDS Society*, vol. 15, no. 3, pp. 1–9.

MARRI (1992). Number of sexual partners in lifetime by marital status and religious attendance. Tech. Rep. 122, Marriage and Religion Research Institute.

Martyn, K.K. and Martin, R. (2003). Adolescent Sexual Risk Assessment. *Journal of Midwifery & Women's Health*, vol. 48, no. 3, pp. 213–219.

May, R.M. and Anderson, R.M. (1987). Transmission dynamics of HIV infection. *Nature*, vol. 326, no. 6109, pp. 137–42.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. 2nd edn. Chapman & Hall, London.

Meyer-Weitz, A., Reddy, P., Ven Den Borne, H., Kok, G. and Pietersen, J. (2003). Determinants of multi-partner behaviour of male patients with sexually transmitted diseases in south africa: implications for interventions. *International Journal of Men's Health*, vol. 2, no. 2, p. 149.

Mishra, V., Hong, R., Assche, S. and Barrere, B. (2009). The role of partner reduction and faithfulness in HIV prevention in sub-Saharan Africa: evidence from Cameroon, Rwanda, Uganda, and Zimbabwe. Tech. Rep. 61, USAID. DHS Working Papers.

Mlambo, M.G., Peltzer, K. and Chirinda, W. (2016). Predictors of multiple concurrent and multiple sexual partnerships among male and female youth aged 18–24 in south africa. *Journal of Psychology in Africa*, vol. 26, no. 2, pp. 156–163.

Mongwe, R. (2002). Migration study in the Western Cape 2001: Residents' perceptions regarding migration and social service delivery (especially health and education): Case studies in George and Cape Town. Tech. Rep., Provincial Government of the Western Cape.

Navarro, P., Bekker, L.G., Darkoh, E. and Hecht, R. (2010). Special report on the state of HIV/AIDS in South Africa. *Global Health*, vol. 3, pp. 1–8.

NE, M., GA, W., WA, F. and et al (1990). High-risk std/hiv behavior among college students. *JAMA*, vol. 263, no. 23, pp. 3155–3159. /data/journals/jama/9258/jama_263_23_031.pdf. Available at: +http://dx.doi.org/10.1001/jama.1990.03440230051031

NIH (2015). Drug and alcohol use – a significant risk factor for HIV. Tech. Rep., National Institute on Drug Abuse.

No, C.M. (2002). Hiv risk factors: A review of the demographic, socio-economic, biomedical and behavioural determinants of hiv prevalence in south africa. Tech. Rep. 8.

Odimegwu, C. and Somefun, O.D. (2017). Ethnicity, gender and risky sexual behaviour among nigerian youth: an alternative explanation. *Reproductive health*, vol. 14, no. 1, p. 16.

Otutubikey Izugbara, C. and Nwabuawele Modo, F. (2007). Risks and benefits of multiple sexual partnerships: beliefs of rural nigerian adolescent males. *American journal of men's health*, vol. 1, no. 3, pp. 197–207.

Park, A. (2012). Truvada: 5 things to know about the first drug to prevent HIV. Accessed: November 28, 2016).
Available at: http://healthland.time.com/2012/07/17/truvada-5-things-to-know-about

Parker, W., Makhubele, B., Ntlabati, P. and Connolly, C. (2007). Concurrent sexual partnerships amongst young adults in south africa. challenges for hiv prevention communication.

Patience, O. and Osagie, M. (2014). Modeling the prevalence of malaria in Niger State : An application of Poisson regression and negative binomial regression models. vol. 2, no. 4, pp. 61–68.

PHAC (2012). HIV transmission risk: a summary of the evidence. Tech. Rep., Public Health Agency of Canada.

Plunkettfor, M. (2014). Sex around the world in 10 facts. Accessed: November 03, 2016.
Available at: http://www.best-country.com/article/10-Facts

Quin, C., Mann, M., Curran, W. and Piot, P. (1986). Aids in africa: an epidemiological paradigm. *Science*, vol. 234, pp. 955–63.

Randolph, M.E., Pinkerton, S.D., Bogart, L.M., Cecil, H. and Abramson, P.R. (2007). Sexual pleasure and condom use. *Archives of Sexual Behavior*, vol. 36, no. 6, pp. 844–848.

Reniers, G. and Watkins, S. (2010). Polygyny and the spread of HIV in sub-Saharan Africa: a case of benign concurrency. *AIDS (London, England)*, vol. 24, no. 2, pp. 299–307.

Roberts, P., Holmes, K., Piot, P., Plummer, F., D'Costa, L., Lightfoote, M., Koech, D., Kreiss, J. and Ronald, A. (1986). AIDS virus infection in Nairobi prostitutes - spread of the epidemic to East Africa. *The New England Journal of Medicine*.

Rodrıguez, G. (2001). Smoothing and non-parametric regression. Working paper.

Roper, W.L., Peterson, H.B. and Curran, J.W. (1993). Commentary: condoms and HIV/STD prevention–clarifying the message. *American Journal of Public Health*, vol. 83, no. 4, pp. 501–503.

Rosenbaum, E. and Kandel, D.B. (1990). Early onset of adolescent sexual behavior and drug involvement. *Journal of Marriage and the Family*, vol. 52, no. 3, pp. 783–798.

Santelli, J.S., Brener, N.D., Lowry, R., Bhatt, A. and Zabin, L.S. (1998). Multiple sexual partners among U.S. adolescents and young adults. *Family Planning Perspectives*, vol. 30, no. 6, pp. 271–275.

Sathiyasusuman, A. *et al.* (2015). Associated risk factors of stis and multiple sexual relationships among youths in malawi. *PloS one*, vol. 10, no. 8, p. e0134286.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. 3rd edn. Cambridge university press, New York.

SDI and GIS (2013). City of Cape Town – 2011 census suburb Delft. Tech. Rep., Compiled by Strategic Development Information and GIS Department (SDI and GIS).

SDI and GIS (2013). City of Cape Town – 2011 Census suburb overview Constantia. Tech. Rep., Strategic Development Information and GIS Department.

Seekings, J. (2013). Economy, society and municipal services in Khayelitsha. *Report for the Commission of Inquiry into Allegations of Police Inefficiency in Khayelitsha and a breakdown in relations between the community and the police in Khayelitsha, Centre for Social Science Research, University of Cape Town*.

Shikin E.V. and Plis A.I. (1995). *Handbook on splines for the User*. CRC Press, London.

Shisana, O., Rehle, T., Simbayi, L., Parker, W., Zuma, K., Bhana, A., Connolly, C., Jooste, S. and Pillay, V. (2005). *South African national HIV prevalence, HIV incidence, behaviour and communication survey, 2005*.

Shisana, O., Rehle, T., Simbayi, L., Zuma, K., Jooste, S., Zungu, N., Labadarios, D. and Onoya, D. (2012). South African National HIV Prevalence, Incidence and Behaviour Survey, 2012. p. 194.

Smith, T.W. (1992). Discrepancies between men and women in reporting number of sexual partners: A summary from four countries. *Social Biology*, vol. 39, no. 3-4, pp. 203–211.

Stigum, H., Magnus, P., Harris, J.R., Samuelsen, S.O. and Bakketeig, L.S. (1997). Frequency of sexual partner change in a Norwegian population. Data distribution and covariates. *Am J Epidemiol*, vol. 145, no. 7, pp. 636–643.

Suttinee, K. (2002). Importance of models in economics. Tech. Rep..

Tanfer, K. and Schoorl, J.J. (1992). Premarital sexual careers and partner change. *Archives of Sexual Behaviour*, vol. 21, no. 1, pp. 45–68.

Temple-Smith, M. (2014). *Sexual Health: A multidisciplinary approach*. IP Communications, Pty. Ltd., Australia.

Thobejane, T.D. and Flora, T. (2014). An exploration of polygamous marriages: A worldview. *Mediterranean Journal of Social Sciences*, vol. 5, no. 27 P2, p. 1058.

Todd, J., Cremin, I., McGrath, N., Bwanika, J., Wringe, A., Marston, M., Kasamba, I., Mushati, P., Lutalo, T., Hosegood, V. and Zaba, B. (2009). Reported number of sexual partners: comparison of data from four African longitudinal studies. *Sexually transmitted infections*, vol. 85 Suppl 1, pp. i72–80.

Tsatsou, M. (2012). Durex survey: Greeks have the most sex weekly but nigerians are most satisfied. Accessed: November 03, 2016.
Available at: http://greece.greekreporter.com/2012/07/19/durex-survey-greeks-have-the-most-sex-weekly/

Tutz, G. (2011). *Regression for categorical data*, vol. 34. Cambridge University Press.

UNAIDS (1998). Looking deeper into the HIV epidemic. Tech. Rep., Joint United Nations Programme on HIV/AIDS.

UNAIDS (2013). Global report: UNAIDS report on the global AIDS epidemic 2013. Tech. Rep., Joint United Nations Programme on HIV/AIDS.

UNAIDS (2014). South Africa: HIV and AIDS estimates. Accessed: November 17, 2017).
Available at: http://www.unaids.org/en/regionscountries/countries/southafrica/

UNAIDS (2016). Fact sheet 2016 global statistics. Tech. Rep., Joint United Nations Programme on HIV/AIDS.

UNAIDS, UNESCO and Commission of the European Communities (2000). Migrant populations and HIV/AIDS : the development and implementation of programmes : theory, methodology and practice. Tech. Rep., Joint United Nations Programme on HIV/AIDS. UNAIDS best practice collection key material.

Wacholder, S. (1986). Binomial regression in GLIM: Estimating risk ratios and risk differences. *American Journal of Epidemiology*, vol. 123, no. 1, p. 174.

Watts, C.H. and May, R.M. (1992). The influence of concurrent partnerships on the dynamics of hiv/aids. *Mathematical biosciences*, vol. 108, no. 1, pp. 89–104.

Wellings, K., Collumbien, M., Slaymaker, E., Singh, S., Hodges, Z., Patel, D. and Bajos, N. (2006). Sexual behaviour in context: a global perspective. *Lancet*, vol. 368, no. 9548, pp. 1706–1728.

Wellings, K. and Mitchell, K. (2012). *Sexual health: a public health perspective*. McGraw-Hill Education (UK).

WHO (2007). Information package on male circumcision and HIV prevention. Tech. Rep. Insert 4, World Health Organization, UNAIDS, UNICEF, UNFPA, The World Bank.

Wilton, B. (2014). Getting to the bottom of it: anal sex, rectal fluid and hiv transmission. *Toronto, Canada: CATIE*.

Winkelmann, R. (2013). *Econometric analysis of count data*. Springer Science & Business Media, Germany.

Winkelstein, W., Lyman, D.M., Padian, N., Grant, R., Samuel, M., Wiley, J.A., Anderson, R.E., Lang, W., Riggs, J. and Levy, J.A. (1987). Sexual practices and risk of infection by the human immunodeficiency virus: the san francisco men's health study. *The Journal of the American Medical Association*, vol. 257, no. 3, pp. 321–325.

Wittrock, L.a. (2004). The Gender Discrepancy in Reported Number of Sexual Partners : Effects of Anonymity. *Psychology*, pp. 1–5.

Wold, S. (1974). Spline functions in data analysis. *Technometrics*, vol. 16, no. 1, pp. 1–11.

Zimmer-Gembeck, M. and Collins, W. (2008). Sexual partnering - ages 16 to 26. *The Journal of Adolescent Health*, pp. 564–572.

Zou, G. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, vol. 159, no. 7, pp. 702–706.

Zou, G. and Donner, A. (2013). Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Statistical Methods in Medical Research*, vol. 22, no. 6, pp. 661–670.